Rui Jorge Viegas Gonçalves Pereira

Licenciado em Matemática

Tarifação *a priori* - Estudo de uma carteira automóvel

Dissertação para obtenção do Grau de Mestre em

Matemática e Aplicações
Ramo Atuariado, Estatística e Investigação Operacional

Orientadora: Gracinda Rita Diogo Guerreiro, Prof. Auxiliar, Universidade Nova de Lisboa

Júri

Presidente:	Manuel Leote Tavares Inglês Esquível
Arguente:	Maria de Lourdes Belchior Afonso
Vogal:	Gracinda Rita Diogo Guerreiro

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2017

Tarifação *a priori* - Estudo de uma carteira automóvel

"O sucesso é ir de fracasso em fracasso sem perder entusiasmo."
– Winston Churchill

Agradecimentos

À Professora Gracinda Guerreiro, orientadora desta dissertação, pela disponibilidade e apoio incondicional ao longo de todo este projeto. Agradeço também a confiança que depositou em mim, desde o início, mas também o sentido de responsabilidade que me incutiu em todas as fases desta dissertação.

À Dr.ª Carmen Oliveira, o meu sincero agradecimento pela oportunidade que me deu de integrar na Actuariado e desenvolver esta dissertação. O seu apoio, acompanhamento e orientação foram muito importantes.

Às minhas colegas Cláudia Conceição e Rute Ferreira, obrigada pela recetividade com que me acolheram, por toda a amizade e companheirismo e ajuda, fatores muito importantes na realização deste trabalho.

À minha namorada, Rita Machado, um agradecimento especial por todas as palavras de incentivo e pela transmissão de confiança e de força, em todos os momentos.

À minha Família, primeiramente aos meus Irmãos, por acreditarem sempre em mim e prestarem-me todo o apoio necessário ao longo de toda a minha vida. Por fim, e sem menos importância, aos meus Pais, por serem as pessoas mais importantes e porque sem eles nada disto seria possível.

A todos eles, os meus maiores agradecimentos!

Resumo

O objetivo desta dissertação consiste na construção de uma tarifa *a priori* para uma carteira de Responsabilidade Civil Automóvel. A elaboração de uma tarifa automóvel clássica passa, fundamentalmente, pela agregação de segurados da carteira em grupos que se considerem como riscos homogéneos – os Escalões Tarifários. Com esta finalidade, serão abordadas duas possíveis metodologias para a construção da tarifa: a primeira, utilizando os modelos lineares generalizados (MLG) e, na segunda, recorrendo à modelação das indemnizações agregadas, através da Família de Distribuições Tweedie.

Os resultados obtidos permitirão avaliar qual a melhor metodologia a seguir para uma tarifação mais adequada à carteira em estudo. Os dados da carteira foram fornecidos por uma Seguradora a operar em Portugal.

Palavras-chave: Tarifa Automóvel, Responsabilidade Civil Automóvel, Tarifa *a priori*, Modelos Lineares Generalizados, Família Exponencial, Família de Distribuições Tweedie.

Abstract

The purpose of this dissertation is to construct an *a priori* tariff for a Third Party Liability Motor Insurance. The elaboration of a car tariff fundamentally involves the aggregation of policyholders into groups that are considered as homogenous risks - the Tariff Scales. For this purpose, it will be discussed two possible methodologies for the construction of the tariff: the first one, using Generalized Linear Models (GLM) and, in the second, using modeling of the aggregate claims, through the Tweedie Family of Distributions.

The results obtained will allow us to evaluate the best methodology for more appropriate pricing for the portfolio under study. The data from the portfolio was provided by a Insurance company operating in Portugal.

Keywords: Car Tariff, Motor Third Party Liability, a priori Tariff, Generalized Linear Models, Exponential Family, Tweedi Family of Distribuition.

Índice

Lista de Figuras

Lista de Tabelas

Siglas

AIC Akaike Information Criterion.

ASF Autoridade de Supervisão de Seguros e Fundos de Pensões.

BIC Bayesian Information Criterion.

MLG Modelos Lineares Generalizados.

RC Responsabilidade Civil.

1

Introdução

O sector segurador em Portugal tem tido nos últimos anos um balanço positivo, tendo em conta a conjuntura económica e financeira em que se vive. É necessário analisar os desafios futuros e compreender qual a situação económica e financeira e como esta vai evoluir, bem como os impactos das mudanças regulatórias nas companhias de seguros, segundo a Autoridade de Supervisão de Seguros e Fundos de Pensões (ASF), ver ASF (2016).

A atividade seguradora é uma atividade de gestão de riscos que tem como principal vantagem, em relação aos outros sectores da economia, proporcionar aos seus clientes a satisfação de uma futura necessidade a um custo justo e equitativo. Para tal, a seguradora recorre a um acordo celebrado entre as duas partes, o contrato de seguro, onde o segurado fica obrigado ao pagamento de um prémio.

Existem dois grandes grupos de seguros, os seguros do Ramo Vida e os do Ramo Não Vida. Ao longo dos últimos dois anos, o ramo vida tem registado uma forte redução de produção de seguro directo, ao contrário do que se tem verificado para o ramo não vida. É no ramo não vida que estão contemplados os seguros automóveis, tendo-se verificado em 2016, o terceiro seguro com maior crescimento, ver ASF (2016).

O seguro automóvel obrigatório por lei, contempla apenas a cobertura de Responsabilidade Civil (RC) que, como já referido, responsabiliza a seguradora pelo pagamento dos danos materiais e/ou corporais causados a terceiros, como descrito no ponto 1 do artigo 4.º do Decreto-lei 291-2007, que estabelece que:

"Toda a pessoa que possa ser civilmente responsável pela reparação de danos corporais ou materiais causados a terceiros por um veículo terrestre a motor para cuja condução seja necessário um título específico e seus reboques, com estacionamento habitual em Portugal, deve, para que esses veículos possam circular, encontrar-se coberta por um seguro que garanta tal responsabilidade, nos termos do presente decreto-lei.".

Uma das técnicas utilizadas para o cálculo do prémio, designada como tarifação *a*

priori, consiste numa técnica de modelação estatística a partir da qual a seguradora tenta prever a frequência e o custo da sinistralidade futura, tendo em consideração as características dos segurados que influenciam o risco inerente ao contrato, tais como, a idade, o tipo de carro, local de residência, entre outros. Através desta técnica, as seguradoras procuram agrupar segurados em grupos de risco homogéneos, de tal forma que todos os segurados pertencentes à mesma classe paguem o mesmo valor de prémio.

O objetivo da presente dissertação passa pela construção de uma tarifa automóvel de Responsabilidade Civil para a Empresa Actuariado, de modo a aprofundar conhecimento na área de tarifação, para que no futuro possam ser utilizadas novas metodologias nesta área. Esta dissertação foi elaborada tendo em conta duas abordagens: Uma abordagem mais clássica para a construção de tarifas *a priori*, baseada nos Modelos Lineares Generalizados, para modelar a frequência de sinistralidade bem como a severidade dos sinistros. A segunda abordagem reside na utilização dos modelos de dispersão desenvolvidos por Jorgensen (1997), onde se insere a distribuição Tweedie, uma distribuição que permite a modelação das indemnizações agregadas, sem realizar a modelação em separado da frequência de sinistralidade e da severidade dos sinistros. Pretende-se apresentar os resultados das duas metodologias não sendo, no entanto, nosso objetivo a escolha de uma tarifa considerada ideal e/ou adequada para comercialização do seguro.

A presente dissertação está dividida em cinco capítulos. O primeiro capitulo – Introdução – é dedicado ao enquadramento e importância do tema escolhido, definição dos objetivos e estrutura da dissertação. No segundo capítulo são abordados conceitos usuais nesta atividade de seguros, tais como o conceito de prémio, os seus princípios de cálculo e a técnica de tarifação *a priori*. São também referidos fatores que influenciam a frequência de sinistralidade. Por fim são apresentados os processos de Poisson homogéneos e mistos, bem como as distribuições usualmente utilizadas na modelação da severidade dos sinistros participados à seguradora.

O terceiro capítulo aborda conceitos estatísticos necessários para o presente estudo, mais concretamente, a evolução e definição dos Modelos Lineares Generalizados e os Modelos de Dispersão, onde se abordam a Família Exponencial e a Família Tweedie. São também descritos métodos para a estimação dos parâmetros dos modelos bem como métodos para aferir a qualidade dos ajustamentos.

O quarto capítulo é dedicado à componente prática da dissertação, iniciando-se por uma breve apresentação da Empresa Actuariado, sendo de seguida apresentados os dados para a construção da tarifa. É neste capítulo que é efetuado o tratamento estatístico, bem como os estudos preliminares efetuados à base de dados. No fim, é apresentada a tarifa obtida através dos Modelos Lineares Generalizados, resultado da modelação das variáveis aleatórias Número e Severidade dos Sinistros, com base em distribuições da Família Exponencial e comparada com a tarifa obtida recorrendo à Família de Distribuições Tweedie.

O último capítulo contém as conclusões que resultam da análise dos resultados obtidos nesta dissertação.

2

Tarifação *a priori*

Um dos grandes desafios das seguradoras centra-se na estimação do prémio de seguro a cobrar por cada contrato celebrado, mediante as coberturas contratadas. No caso particular do seguro automóvel e de outros seguros de massa, onde vários segurados têm características comuns entre si, é frequente que em vez de calcular um prémio individual, se construa uma tarifa, que irá tabular o prémio a cobrar por apólice, de acordo com as suas caraterísticas.

O prémio de seguro cobrado na celebração do contrato fará face a futuras sinistralidades, de montantes que são desconhecidos *a priori*. No caso particular do seguro de Responsabilidade Civil automóvel, o capital mínimo seguro obrigatório é de 7.290.000€, subdividido em 6.070.000€ para danos corporais e 1.220.000€ para danos materiais (valores de 2017), ver artigo 12.º do Decreto-Lei n.º 291/2007, transmitindo a ideia de que, em caso de ocorrência de sinistros mais graves, as respetivas indemnizações a pagar podem ser de valores muitos avultados.

O prémio a pagar pelo segurado deve ser justo, caso contrário, se da tarifa resultarem prémios demasiado elevados, estes podem inibir a entrada de novos segurados e possivelmente a saída de outros, que resultaria na perda de carteira. Já no caso de resultarem prémios abaixo do que seria comercialmente esperado, as responsabilidades futuras da seguradora poderão estar comprometidas.

No decorrer deste capítulo serão apresentadas vários terminologias e conceitos básicos importantes para a construção da tarifa, que serão utilizados ao longo desta dissertação.

2.1 Noções Introdutórias

O contrato de seguro é um acordo realizado entre o segurado e a seguradora. Este assegura a cobertura de determinados riscos por parte da seguradora, ficando esta responsabilizada

por pagar as indemnizações em caso de ocorrência de sinistros, de acordo com os termos acordados entre ambas as partes. Em contrapartida, o segurado fica obrigado a pagar um montante monetário, o prémio.

Do ponto de vista de uma seguradora, uma das suas maiores preocupações baseia-se na determinação do prémio de seguro de forma a que este reflita, tanto quanto possível, o risco associado ao contrato. Neste contexto, o prémio é um montante monetário pago pelo tomador de seguro[1], em contrapartida das garantias que lhe são concedidas pela seguradora, previamente acordadas no contrato.

O prémio puro de um contrato de seguro corresponde à Esperança Matemática do risco a segurar. De uma forma simples, ver Ohlsson e Johansson (2010), pode dizer-se que é calculado através de:

$$\text{Prémio Puro} = \text{Frequência} \times \text{Severidade} \tag{2.1}$$

Analisando a expressão anterior, é notório que o prémio puro tem de ser calculado de forma a fazer face a futuras sinistralidades, tendo assim em conta, a frequência e a severidade estimadas dos futuros sinistros.

2.1.1 Princípios de cálculo do prémio

Um princípio de cálculo de prémio é uma regra (função que se designará por Π) que permite calcular o prémio de seguro. Existem diferentes princípios de cálculo de prémio que conduzem, consequentemente, a valores distintos.

Designe-se por P o prémio de um contrato de seguro, por S o montante das indemnizações agregadas de uma carteira de seguros de um determinado ramo, e por $\alpha > 0$ a carga de segurança, associada ao cálculo do prémio.

$$S = \sum_{i=1}^{N} Y_i \tag{2.2}$$

em que:

- N - número de sinistros da carteira;
- Y_i - Montante do i-ésimo sinistro $i = 1, \cdots, N$.

Note-se que N é uma variável aleatória discreta e Y_i uma variável aleatória contínua. Por conseguinte, S, o montante total das indemnizações originadas, representa um processo composto. Note-se ainda que as variáveis explicativas que influenciam a variável aleatória N, poderão não ser as mesmas que influenciam as variáveis aleatórias Y_i. Por todas estas razões, é comum efetuar-se a modelação estatística do Número e Severidade dos sinistros em separado.

[1] Pessoa que celebra o contrato de seguro

O Prémio Puro corresponde à Esperança Matemática do risco, PP=$\mathbb{E}(S)$, em que $\mathbb{E}(S)$ é o valor que a seguradora espera vir a pagar em indemnizações relativamente a uma determinada carteira e que corresponde ao valor esperado das indemnizações agregadas ver, por exemplo, Centeno (2002).

Para se efetuar a modelação do prémio puro, o objetivo passa então por estimar $\mathbb{E}(S)$.

Considerando que os montantes Y_i, $i = 1, \cdots, N$ são independentes do número de sinistros e que são identicamente distribuídos entre si, tem-se segundo Goovaerts et al. (2001):

$$\mathbb{E}(S) = \mathbb{E}(N) \times \mathbb{E}(Y) = \mathbb{E}\left(\sum_{i=1}^{N} Y_i\right), \quad i = 1, \cdots, N$$

Contudo, para além do princípio de cálculo do prémio anteriormente referido, existem outros princípios que são frequentemente adotados ver, por exemplo, Centeno (2002):

- **Princípio do Valor Esperado:** Π(S)=(1+α)$\mathbb{E}$(S)
- **Princípio da Variância:** Π(S)=$\mathbb{E}$(S)+$\alpha\mathbb{V}(S)$
- **Princípio do Desvio Padrão:** Π(S)=$\mathbb{E}$(S)+$\alpha\sqrt{\mathbb{V}(S)}$

uma vez que visam obter um prémio monetário mais elevado de modo a fazer face a eventuais desvios dos montantes de indemnizações.

2.1.2 Tarifa

Aquando da estimação do prémio a cobrar, e uma vez que este tem por base os dados históricos de sinistralidade, a seguradora terá de ter em consideração o tipo de perfil de cada segurado e o risco que este representa.

Na construção de uma tarifa, e perante uma carteira que contenha riscos com um certo grau de heterogeneidade, é necessário identificar as variáveis (os fatores tarifários) que influenciam esse mesmo risco, criando posteriormente grupos que apresentem riscos semelhantes, ou seja, grupos homogéneos de risco (os escalões tarifários) permitindo, assim, estimar o prémio adequado a cada apólice mediante o escalão em que esta se encontra.

Naturalmente, apólices que se apresentem no mesmo nível tarifário apresentarão prémios semelhantes. A tarifa permite, assim, definir um conjunto de regras que permitem estimar o prémio a pagar por cada apólice.

Enquanto a tarifa se refere ao prémio a cobrar a um segurado com determinadas características, a estrutura tarifária refere-se à relação existente entre os prémios de uma carteira, em função dos fatores que servem de base ao cálculo da tarifa.

O escalão tarifário base representa um conjunto de características associadas a cada uma das variáveis tarifárias, e é usualmente designado como *Segurado Padrão*. Este permite estimar o prémio base da tarifa, a partir do qual se obtém os prémios dos restantes

escalões tarifários, efetuando agravamentos ou descontos, consoante a análise de risco que é efetuada entre o Segurado Padrão e os demais escalões. Assim, os restantes escalões em estudo poderão ter prémios superiores ou inferiores ao prémio base, consoante maior ou menor risco estimado nos restantes escalões face ao Segurado Padrão.

2.2 Tarifa *a priori*

No processo de determinação do montante dos prémios para o conjunto de apólices da carteira, existem duas abordagens que se complementam: a tarifação *a priori* e *a posteriori*.

Tal como referido na secção anterior, a seguradora constrói uma tarifa subdividindo a carteira em grupos de risco homogéneos. Deste modo, é fundamental determinar os fatores relevantes que têm impacto no risco. Caso o mesmo não aconteça, serão agrupadas apólices com riscos de gravidades diferentes, o que poderá originar uma tarifa desadequada.

A técnica de tarifação *a priori*, pode ilustrar-se no pagamento de prémios diferenciados em função de fatores mensuráveis do segurado, como exemplo:

- Anos de carta
- Idade do veículo
- Zona de residência
- Marca do veículo
- Idade segurado
- Quilómetros anuais
- Potência do veículo
- Tipo de combustível

A modelação das estruturas tarifárias *a priori* é frequentemente efetuada com base nos Modelos Lineares Generalizados ver, por exemplo, Ohlsson e Johansson (2010). Estes permitem estimar a influência dos fatores de tarifação sobre a frequência de sinistralidade bem como sobre a severidade dos sinistros.

No cálculo de prémios para os diferentes escalões tarifários, existem dois modelos possíveis, modelo aditivo ou multiplicativo, ver, por exemplo, Ohlsson e Johansson (2010).

Os modelos para M fatores tarifários podem ser apresentados da seguinte forma:

Modelo Multiplicativo: $\mathbb{E}[Y_{i_1,\cdots,i_M}] = \mu_{i_1,\cdots,i_M} = \gamma_0\gamma_{1i_1}\gamma_{2i_2}\cdots\gamma_{Mi_M}$

Modelo Aditivo: $\mathbb{E}[Y_{i_1,\cdots,i_M}] = \mu_{i_1,\cdots,i_M} = \gamma_0 + \gamma_{1i_1} + \gamma_{2i_2} + \cdots + \gamma_{Mi_M}$

No modelo multiplicativo o prémio de cada segurado/apólice é obtido por multiplicação do Prémio do segurado padrão (γ_0) pelos coeficientes associados a cada escalão tarifário $\gamma_{j\,i_j}\ j = 1,\cdots,M$.

Num modelo aditivo, o prémio de cada escalão é obtido a partir do prémio do Segurado Padrão (γ_0) ao qual são posteriormente adicionados ou descontados valores monetários $\gamma_{j\,i_j}\ j = 1,\cdots,M$ consoante as características do segurado que diferem do Segurado Padrão.

Em termos de construção da tarifa, a opção entre do modelo aditivo ou multiplicativo traduzir-se-á na escolha da função de ligação (que será apresentada no próximo capítulo) a utilizar nos Modelos Lineares Generalizados. De um ponto de vista prático é muito mais comum as seguradoras construirem tarifas multiplicativas em detrimento de tarifas aditivas. Algumas justificações relevantes acerca desta opção podem ser consultadas em Ohlsson e Johansson (2010) e Guerreiro (2016).

Tal como visto anteriormente, a frequência de sinistralidade e a severidade podem ser afetadas por inúmeros fatores. Assim, aquando da modelação do prémio puro, a frequência e o custo dos sinistros são modelados, por norma, de forma independente.

2.2.1 Construção da tarifa *a priori*

Uma vez que que as variáveis explicativas podem ter impactos diferentes sobre o custo e sobre a frequência de sinistralidade, é usual proceder-se ao estudo em separado das mesmas. Assim, assumindo a independência entre o número e o custo dos sinistros são apresentadas as metodologias nos pontos seguintes.

2.2.1.1 Modelação da Frequência de sinistralidade

Quando se pretende estudar o número de sinistros de uma carteira do Ramo Automóvel, é comum representar-se este fenómeno através de um processo de Poisson pelo que, no que se segue, se apresentarão alguns resultados relevantes acerca deste tipo de processos de contagem.

Processo de Poisson Homogéneo

Seja $N(t)$ o número de indemnizações (ou número de sinistros), relativas a um determinado risco, ocorridas num intervalo de tempo $[0,t[$, com $t > 0$ e $N(t) = 0$.

Suponha-se que $\{N(t)\}_{t>0}$ é um processo de contagem. Diz-se que $\{N(t)\}_{t>0}$ é um processo de Poisson homogéneo de intensidade λ, se verificar as seguintes condições ver, por exemplo, Ross (1996):

1. $\{N(t)\}_{t>0}$ tem incrementos independentes;

2. $\{N(t)\}_{t>0}$ tem incrementos estacionários;

3. Para $\forall h \longrightarrow 0^+$, $P_r\{N(h) \geq 1\} = \lambda h + o(h)$;

4. Para $\forall h \longrightarrow 0^+$, $P_r\{N(h) \geq 2\} = o(h)$.

em que $o(h)$ é um infinitésimo. Recorde-se que uma função f é um infinitésimo, representando-se por $o(h)$ quando

$$\lim_{h\to 0} \frac{f(h)}{h} = 0$$

Satisfazendo as condições anteriores, $\{N(t)\}_{t>0}$ é um processo de Poisson Homogéneo. Desta forma, para cada instante t, a variável aleatória $N(t)$ tem distribuição de Poisson de média λt, com $t > 0$, isto é

$$p_x(t) = P[N(t) = x] = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x \epsilon \mathbb{N}_0, \quad \lambda > 0 \tag{2.3}$$

onde λ é designado por intensidade do processo e representa o número de sinistros ocorridos por unidade de tempo.

Dado que se está perante uma distribuição de Poisson, o valor médio e a variância da distribuição do número de sinistros no presente intervalo de amplitude t, são iguais a λt.

Uma vez que cada segurado tem características diferentes, ou seja, um segurado pode ter mais propensão para causar acidentes que outro, é usual que, numa carteira, o parâmetro λ não seja o mesmo para todos eles.

Consequentemente, os processos de Poisson Homogéneos revelam-se desadequados para estimar a frequência de sinistralidade. Assim, justifica-se que se suponha que o número de sinistros possa ter uma distribuição de Poisson em que o parâmetro λ resulte da observação de uma variável aleatória, Λ, não negativa.

Processo de Poisson Heterogéneo

Nos processos de Poisson Heterogéneos, ou Mistos, considera-se que a intensidade do processo λ, varia de segurado para segurado, ou seja, considera-se que o número de sinistros segue um processo de Poisson em que o parâmetro é proveniente da observação de uma variável aleatória não negativa, Λ.

Seja

$$U(\lambda) = P[\Lambda \leq \lambda]$$

a função de distribuição da variável aleatória Λ. $U(\lambda)$ e Λ são designadas por distribuição e variável de estrutura, respetivamente.

Para modelar o número de sinistros, recorre-se às distribuições mais usuais, a distribuição Poisson e Binomial Negativa, estimando-se os respetivos parâmetros. Esta modelação tem por base um grande número de observações, considerando N_i o número de sinistros por apólice i, em que $i = 1, \cdots, n$.

O processo de contagem $\{N(t)\}_{t>0}\}$ designa-se por Processo de Poisson Misto se satifazer as seguintes condições, ver, por exemplo, Centeno (2002):

1. $N(0) = 0$;
2. $P[N(t+s) - N(s) = x] = \int_0^{+\infty} P[N(t+s) - N(s) = x|\Lambda = \lambda] = \int_0^{+\infty} e^{-\lambda t}\frac{(\lambda t)^x}{x!} dU(\lambda)$

Tendo-se que a média e a variância da distribuição de Poisson mista são dadas por:

$$\mathbb{E}[N(t)] = t\mathbb{E}[\Lambda] \qquad \text{e} \qquad \mathbb{V}[N(t)] = t\mathbb{V}[\Lambda] + t^2\mathbb{V}[\Lambda]$$

Um caso particular do processo de Poisson Misto é o modelo Poisson-Gama, ou processo de Polya, ver, por exemplo, Grandell (1997), quando a variável Λ segue um distribuição Gama(α, β), ou seja, quando a sua função densidade é da forma:

$$u(\lambda) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} e^{-\beta\lambda} \lambda^{\alpha-1}, \quad \lambda, \alpha, \beta > 0 \tag{2.4}$$

Por conseguinte, o valor esperado e a variância vêm dados por:

$$\mathbb{E}[\Lambda] = \frac{\alpha}{\beta} \qquad \text{e} \qquad \mathbb{V}[\Lambda] = \frac{\alpha}{\beta^2}$$

Assim, se

$$N(t) \sim P(\Lambda t) \qquad \text{com} \qquad \Lambda \sim G(\alpha, \beta)$$

vem que, ver, por exemplo, Centeno (2002),

$$N(t) \sim BN\left(\alpha, \frac{\beta}{\beta + t}\right)$$

Se se considerar como unidade temporal um ano, ou seja, $t = 1$, obtém-se, para $N \equiv N(1)$,

$$\mathbb{E}[N] = \frac{\alpha}{\beta} \qquad \text{e} \qquad \mathbb{E}[N] = \frac{\alpha}{\beta}\left(\frac{\beta + 1}{\beta}\right)$$

2.2.1.2 Modelação da Severidade dos Sinistros

Com a finalidade de modelar o custo dos sinistros de uma indemnização Y_i, $i = 1, \cdots, N$, e uma vez que as indemnizações são valores reais positivos, as distribuições mais frequentes, ver por exemplo, Frees et al. (2016b) ou Achieng (2010), são a Gama, a Lognormal ou a Inversa Gaussiana.

Distribuição Gama

Uma variável aleatória Y tem distribuição Gama com parâmetros α e β, escrevendo-se $Y \sim$Gama(α,β) se a sua função densidade de probabilidade é dada por:

$$f(y) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} e^{-\beta y} y^{\alpha-1}, \quad y \geq 0, \alpha, \beta > 0 \tag{2.5}$$

sendo $\mathbb{E}[Y] = \frac{\alpha}{\beta}$ e $\mathbb{V}[Y] = \frac{\alpha}{\beta^2}$.

A distribuição Gama tem suporte em $\mathbb{R}_0^+$. Sendo uma variável aleatória com função densidade enviesada à direita, permite a existência de valores bastante elevados na cauda direita sendo, por isso, uma distribuição bastante utilizada na modelação da severidade dos sinistros. Na Figura 2.1, são ilustrados alguns exemplos da distribuição Gama.

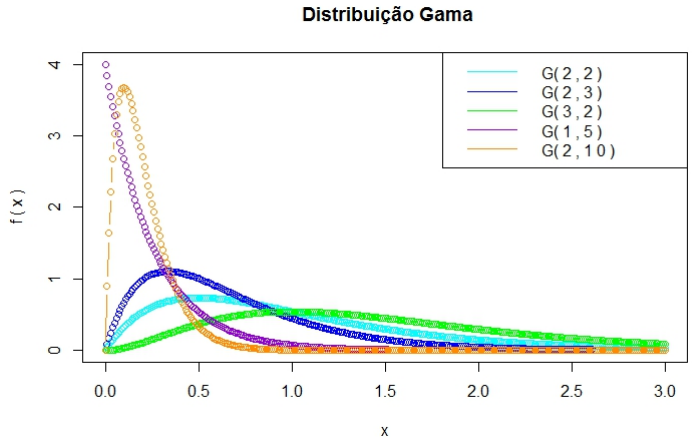

Figura 2.1: Distribuição Gama

Distribuição Inversa Gaussiana

Sendo Y uma v.a que segue uma distribuição Inversa Gaussiana, $Y \sim \text{IG}(\mu,\sigma^2)$, a sua função densidade de probabilidade é dada por

$$f(y) = \frac{1}{\sigma\sqrt{2\pi y^3}} \exp\left\{ -\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2 \right\}, \quad y \in \mathbb{R}_0^+, \mu, \sigma \in \mathbb{R} \tag{2.6}$$

em que $\mathbb{E}[Y] = \mu$ e $\mathbb{V}[Y] = \sigma^2\mu^3$.

Na Figura 2.2, estão presentes exemplos da distribuição Inversa Gaussiana.

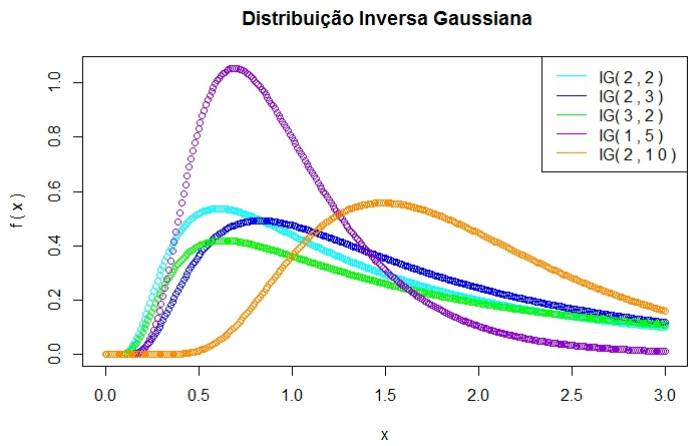

Figura 2.2: Distribuição Inversa Gaussiana

Distribuição LogNormal

Uma variável aleatória Y tem distribuição Lognormal com parâmetros μ e σ^2, escrevendo-se $Y \sim \text{LN}(\mu,\sigma^2)$ quando $Z = \log(Y)$ segue uma distribuição Normal, sendo a sua função densidade de probabilidade dada por

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(\ln(y)-\mu)^2}{2\sigma^2} \right\}, \quad y \in \mathbb{R}_0^+, \mu, \sigma \in \mathbb{R} \tag{2.7}$$

sendo $\mathbb{E}[Y] = e^{\mu+0.5\sigma^2}$ e $\mathbb{V}[Y] = e^{(2\mu+\sigma^2)}\left(e^{\sigma^2}-1\right)$.

Na Figura 2.3, estão presentes exemplos da distribuição Lognormal, considerando diferentes parâmetros.

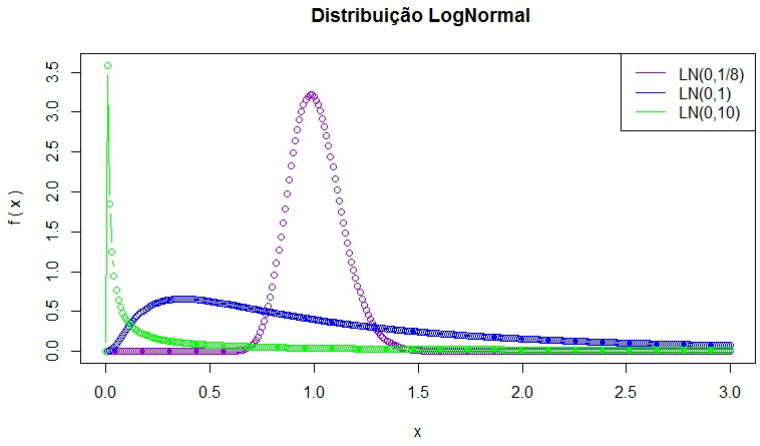

Figura 2.3: Distribuição LogNormal

Fazendo novamente referência à independência entre o número de sinistros e os custos a estes associados, recorde-se que o Prémio Puro de um contrato de seguro corresponde à Esperança Matemática do risco e obtém-se através do produto dos valores esperados do número de sinistros e da severidade dos sinistros:

$$PP = \mathbb{E}[N]\mathbb{E}[Y]$$

2.2.1.3 Grandes Sinistros

Em carteiras de seguro automóvel ocorrem, por vezes, alguns sinistros de carácter excecional mas que representam uma grande percentagem do custo total com sinistros.

Estes sinistros podem condicionar as estimativas, uma vez que, normalmente não seguem a mesma distribuição que os sinistros "usuais". De modo a ultrapassar este problema, é comum separar estes sinistros dos demais e, posteriormente, realizar um estudo de forma a obter resultados que melhor se adequem à realidade da carteira.

Uma das possíveis abordagens, ver, por exemplo, Charpentier (2014), consiste em definir um valor "s" positivo, a partir do qual os sinistros são considerados como "grandes"

sinistros, ou seja:

- Sinistros "usuais" - $Y < s$;
- Sinistros "grandes" - $Y \geq s$.

Tendo em conta resultados conhecidos acerca da Esperança Condicional, ver por exemplo Ross (2014), tem-se:

$$\mathbb{E}[Y] = \mathbb{E}[Y|Y \leq s]\mathbb{P}[Y \leq s] + \mathbb{E}[Y|Y \geq s]\mathbb{P}[Y \geq s] \tag{2.8}$$

o que permitirá modelar separadamente os sinistros considerados "usuais" e os "grandes" sinistros.

Considerando os fatores tarifários $\mathbf{X}$ para a modelação da severidade dos sinistros, a equação (2.8) pode ser substituída por, ver, por exemplo, Charpentier (2014):

$$\mathbb{E}[Y|\mathbf{X}] = \underbrace{\mathbb{E}[Y|\mathbf{X}, Y < s]}_{A}\underbrace{\mathbb{P}[Y < s|\mathbf{X}]}_{C_1} + \underbrace{\mathbb{E}[Y|\mathbf{X}, Y \geq s]}_{B}\underbrace{\mathbb{P}[Y \geq s|\mathbf{X}]}_{C_2} \tag{2.9}$$

em que

- A corresponde ao custo de sinistros "usuais";
- B corresponde ao custo de "grandes" sinistros ;
- C_1 é a probabilidade de ocorrer um sinistro com custo "usual";
- C_2 é a probabilidade de ocorrer um sinistro com custo "grande".

Outra possivel abordagem, apresentada por Ohlsson e Johansson (2010), baseia-se na ideia de que a ocorrência de um "grande" sinistro não é influenciada pelos fatores de tarifação e, como tal, as características observáveis *a priori* para cada segurado dificilmente permitirão prever/estimar a ocorrência/magnitude de um grande sinistro.

Um método usual de modelar estes sinistros é truncar os sinistros, isto é, limitar os montantes das indemnizações a um valor s positivo, tal que valores superiores a este s são considerados como "grandes" sinistros.

Sendo Y_i o custo de um sinistro, considere-se $\widehat{Y_i}$ tal que

$$\widehat{Y_i} = \min(Y_i, s), \quad i = 1, \cdots, n$$

Os autores propõem que na modelação da severidade dos sinistros, se utilize $\widehat{Y_i}$, o que equivale à modelação dos sinistros "usuais", com base nas características de cada escalão tarifário.

Esta abordagem, apresenta dois problemas: qual o melhor valor a escolher para s e como proceder com os valores dos custos que foram truncados.

Usualmente, o valor de s é obtido por tentativa e erro. Por um lado, este tem de ser grande o suficiente para a modelação e as respetivas análises se tornem relevantes e por outro, s terá de assumir um valor pequeno para que as estimativas seja fiáveis.

Nesta técnica, Ohlsson e Johansson (2010) afirmam que uma solução para os custos superiores ao valor truncado s, reside na distribuição uniforme da soma dos mesmos pelos segurados da carteira, contemplando assim o risco de cada segurado originar um "grande" sinistro.

3

Modelos Lineares Generalizados

No início dos anos 70, Nelder e Wedderburn introduzem pela primeira vez os Modelos Lineares Generalizados (MLG). Estes modelos são caraterizados por terem uma estrutura generalizada e também versátil, são também aplicáveis a problemas em que é necessário modelar a relação entre variáveis e estudar a influência destas sobre uma outra variável, permitindo ainda o cálculo do erro presente nas diferentes estimativas.

Os MLG são uma extensão do modelo de regressão linear, onde a distribuição da variável resposta não tem necessariamente que seguir uma distribuição Normal, mas sim uma distribuição pertencente à Família Exponencial de distribuições. Para além disso, a função que relaciona o valor esperado da variável aleatória de interesse e o vetor de variáveis explicativas pode ser qualquer função diferencial, ver por exemplo, Anderson et al. (2004).

Segundo Turkman e Silva (2000), esta metodologia tem vindo a desempenhar um papel fundamental na análise estatística, devido ao grande número de modelos que englobam e à facilidade de análise associada ao rápido desenvolvimento computacional das últimas décadas.

3.1 Evolução dos Modelos Lineares Clássicos

O objetivo dos modelos lineares é, expressar a relação entre uma variável resposta, Y, e um número de variáveis explicativas, também denominadas covariáveis, que são apresentadas num vetor $x = (x_1, \cdots, x_p)^T$. A variável resposta Y pode ser de natureza contínua, discreta ou dicotómica, enquanto as covariáveis podem ser contínuas, discretas, quantitativas de natureza ordinal ou dicotómicas, ver Turkman e Silva (2000).

Assim, este tipo de modelos podem ser descritos através da seguinte expressão:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.1}$$

em que $\mathbf{Y}=(Y_1,\cdots,Y_n)^T$ representa um vetor de dimensões $n\times 1$, de dados observados; $\mathbf{X}$ corresponde à matriz de covariáveis de dimensão $n\times p$, que está usualmente associada a um vetor $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^T$ de parâmetros e $\boldsymbol{\varepsilon}$ é um vetor de erros aleatórios que seguem uma distribuição Normal.

Estes modelos clássicos apresentam alguns pressupostos, tais como:

1. $\mathbb{E}[\varepsilon_i]=0$ e $\mathbb{V}[\varepsilon_i]=\sigma^2$, $\quad i=1,\cdots,n$;
2. As observações $Y_1,\cdots,Y_n$ são independentes;
3. $Y_i\sim N(\mu_i,\sigma^2)$, $\quad i=1,\cdots,n$;
4. $\mu_i=\mathbb{E}[Y_i]=\beta_0+\sum_{j=1}^{p}x_{ij}\,\beta_j$, $\quad i=1,...,n$.

Em 1972, Nelder e Wedderbun propõem os Modelos Lineares Generalizados, apresentando-os como uma extensão dos modelos de regressão linear, ver Nelder e Wedderburn (1972).

OS MLG ver, por exemplo, McCullagh (1984) caracterizam-se por serem constituídos por três componentes, sendo elas:

- **Componente Aleatória** - As componentes de $\boldsymbol{Y}$ são independentes entre si, podendo assumir qualquer distribuição pertencente à Família Exponencial (distribuição Normal, Gama, Poisson, etc.).

- **Componente Sistemática** - As variáveis explicativas passam a ser escritas como combinação linear dada por η_i, o preditor linear.

$$\eta_i=\sum_{j=1}^{p}x_{ij}\beta_j\,,\quad i=1,...,n$$

- **Função Ligação** - A função de ligação para relacionar o valor esperado de Y_i com o preditor linear η_i é uma função $h(\cdot)$ que é diferenciável e monótona, tal que

$$\mu_i=\mathbb{E}[Y_i]=h^{-1}(\eta_i)$$

Esta evolução dos modelos de regressão revela-se muito útil na modelação de estruturas tarifárias, uma vez que as variáveis aleatórias de interesse (número e custo dos sinistros) dificilmente seguem os pressupostos de 1. a 4. dos modelos lineares simples.

3.2 Família Exponencial

Existem, na literatura, inúmeros trabalhos alusivos à Família Exponencial. Nesta secção será utilizado como referência Anderson et al. (2004) e Sen e Singer (1994). Como referido na secção anterior, no contexto dos MLG, pressupõe-se que a variável resposta tenha uma distribuição pertencente à Família Exponencial, ou seja, que a sua função de densidade possa ser escrita na forma:

$$f(\mathrm{y};\theta,\phi) = \exp\Big[\frac{\mathrm{y}\theta - b(\theta)}{a(\phi)} + c(\mathrm{y},\phi)\Big], \quad \mathrm{y} \geq 0, \quad \text{e} \quad \theta,\phi > 0 \tag{3.2}$$

onde:

1. θ é denominado como parâmetro natural ou canónico, relacionado com a média e que pode ou não ser conhecido;
2. ϕ é o parâmetro de dispersão que está relacionado com a variância e é geralmente conhecido;
3. $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas, em que:
   - $a(\phi)$ é uma função positiva e contínua;
   - $b(\theta)$ é diferenciável até à segunda ordem, sendo $b''(\theta) > 0$;
   - $c(\mathrm{y},\phi)$ é independente do parâmetro canónico.

Como referido nos pontos 1. e 2. os parâmetros θ e ϕ representam a informação da média e da variância sobre Y, tal que:

$$\mathbb{E}[Y] = \mu = b'(\theta) \tag{3.3}$$

$$\mathbb{V}[Y] = b''(\theta) \cdot a(\phi) \tag{3.4}$$

Na equação (3.3), é definido θ como uma função de μ. Se for conhecida uma expressão para a função inversa de $b'(\theta)$, então a primeira equação pode ser resolvida para expressar o parâmetro canónico θ em função da média, ou seja:

$$\theta = (b')^{-1}(\mu)$$

Em relação à equação (3.4), esta afirma que a variância de Y é uma função do parâmetro canónico, ou seja, função da média, multiplicada por $a(\phi)$. Muitos autores consideram que $b''(\theta) = \mathbb{V}[\mu]$, assim sendo, vem que a equação (3.4) pode ser reformulada, escrevendo-se:

$$\mathbb{V}[Y] = \mathbb{V}[\mu] \cdot a(\phi)$$

3.2.1 Exemplo - A distribuição Gama

Considere-se que Y$\sim$ $Gama(\alpha,\beta)$ com função densidade dada por:

$$f(\mathrm{y};\alpha,\beta)=\frac{1}{\Gamma(\alpha)}\beta^{\alpha}e^{-\beta \mathrm{y}}\mathrm{y}^{\alpha-1}\,,\qquad y\geq 0,\alpha,\beta>0$$

Considerando a substituição $t=\dfrac{\alpha}{\beta}\Leftrightarrow\beta=\dfrac{\alpha}{\mathrm{t}}$, tem-se:

$$f(\mathrm{y};\alpha,t)=\frac{1}{\Gamma(\alpha)}\Big(\frac{\alpha}{\mathrm{t}}\Big)^{\alpha}e^{-\frac{\alpha}{\mathrm{t}}\mathrm{y}}\mathrm{y}^{\alpha-1}=\exp\left[\frac{-\frac{1}{t}\mathrm{y}-\log(t)}{1/\alpha}+(\alpha-1)\log(\mathrm{y})-\log(\Gamma(\alpha))+\alpha\log(\alpha)\right]$$

Pode assim verificar-se que a função de densidade da distribuição Gama verifica a expressão (3.2), tomando:

$$\theta=-\frac{1}{t}=-\frac{\beta}{\alpha}\,,\qquad \phi=\alpha\,,\qquad b(\theta)=-\log(-\theta)\,,\qquad a(\phi)=\frac{1}{\phi}$$

e

$$c(\mathrm{y},\phi)=(\alpha-1)\log(\mathrm{y})-\log(\Gamma(\alpha))+\alpha\log(\alpha)$$

pelo que se pode afirmar que a distribuição Gama pertence à Família Exponencial.

Por conseguinte, tendo em conta as expressões (3.2) e (3.3), a média e variância de Y são, respetivamente:

$$\mathbb{E}[Y]=\frac{\alpha}{\beta}\quad\text{e}\quad\mathbb{V}[Y]=\frac{\alpha}{\beta^2}$$

A título ilustrativo são seguidamente apresentados mais alguns exemplos de distribuições de funções da Família Exponencial, bem como os seus respetivos parâmetros.

Tabela 3.1: Exemplo de distribuições da família exponencial e respectivos parâmetros

	$N(\mu,\sigma)$	$B(M,\pi)$	$P(\lambda)$	$Gama(\alpha,\beta)$
θ	μ	$\log\left(\frac{\pi}{1-\pi}\right)$	$\log(\lambda)$	β/α
ϕ	σ^2	1	1	α
b(θ)	$\theta^2/2$	$\log(e^{\theta}+1)$	e^{θ}	-log(-θ)
a(ϕ)	ϕ	ϕ	ϕ	$1/\phi$
c(y,ϕ)	$\frac{1}{2}\log(2\pi\sigma^2)$-$\frac{y^2}{2\sigma^2}$	$\log\binom{M}{y}$	- log(y!)	$(\alpha-1)\log(y)-\log(\Gamma(\alpha))+\alpha\log(\alpha)$

Fonte: Adaptado de McCullagh (1984)

3.3 Família de Distribuições Tweedie

A metodologia usual para a construção de uma tarifa automóvel visa identificar a distribuição que melhor se ajusta ao número e à severidade dos sinistros, sendo assim necessário efetuar dois estudos em separado.

Uma alternativa a esta metodologia, ver, por exemplo, Ohlsson e Johansson (2010), incide sobre a Família de Distribuições Tweedie, metodologia esta que efetua a modelação em simultâneo do número e da severidade dos sinistros, efetuando a modelação sobre os montantes agregados dos sinistros da carteira.

A distribuição Tweedie foi estudada inicialmente por Maurice Tweedie, ver Tweedie (1984), mas são Jørgensen e Souza que aprofundam os estudos e as aplicações da mesma, ver por exemplo, Jørgensen e Paes De Souza (1994).

Uma vez que a variável aleatória S, o montante agregado de perda originada por uma dada carteira, ver equação (2.2), apresenta uma grande quantidade de observações nulas, que representam o número de apólices que não participaram sinistros, e contém a componente contínua positiva que representam os valores dos custos dos sinistros é usual recorrer-se à Família de Distribuições Tweedie para efetuar a modelação da variável aleatória S.

Uma variável aleatória Y pertence à família de distribuições Tweedie se a sua variância for escrita na forma, ver demonstração em Jorgensen (1997):

$$\mathbb{V}[Y] = a(\phi)\,\mu^p, \tag{3.5}$$

em que $a(\phi)$ representa uma função positiva e contínua como indicado na descrição da expressão 3.2, μ representa a média da distribuição e p é um parâmetro com domínio em $\mathbb{R}$.

Na Tabela 3.2 são ilustrados os valores do parâmetro p e as respetivas distribuições associadas.

Tabela 3.2: Distribuições da Família Tweedie

Valor de p	Tipo	Distribuição
$p<0$	Contínua	-
$p=0$	Contínua	Normal
$0<p<1$	-	-
$p=1$	Discreta	Poisson
$1<p<2$	Mista não negativa	Poisson Composta
$p=2$	Contínua positiva	Gama
$2<p<3$	Contínua positiva	-
$p=3$	Contínua positiva	Inversa Gaussiana
$p>3$	Contínua positiva	-

Fonte: Adaptado de Ohlsson e Johansson (2010)

A representação da função $b(\theta)$, ver equação (3.2), está também dependente do valor de p:

$$b(\theta) = \begin{cases} e^{\theta}, & \text{se} \quad p = 1; \\ -\log(-\theta), & \text{se} \quad p = 2; \\ -\dfrac{1}{p-2}[-(p-1)\theta]^{(p-2)/(p-1)} & \text{se} \quad 1 < p < 2 \quad \text{ou} \quad p > 2. \end{cases}$$

A primeira derivada de $b(\theta)$ é dada por,

$$b'(\theta) = \begin{cases} e^{\theta}, & \text{se} \quad p = 1; \\ [-(p-1)\theta]^{-1/(p-1)} & \text{se} \quad p > 1. \end{cases}$$

Por conseguinte, através da expressão (3.3) é possível determinar o valor esperado da variável aleatória Y.

Por último, a função de ligação canónica pode ser obtida através da expressão:

$$h(\mu) = \begin{cases} \log(\mu), & \text{se} \quad p = 1; \\ -\dfrac{1}{(p-1)\mu^{p-1}} & \text{se} \quad p > 1. \end{cases}$$

Analisando a Tabela 3.2, pode verificar-se que a Família de Distribuições Tweedie apresenta ser flexível, permitindo a aproximação a várias distribuições da Família Exponencial consoante o valor de p e efetuar a modelação das indemnizações agregadas que, vimos anteriormente, são usualmente modeladas através de Processos de Poisson Compostos.

Apesar da diversidade de aplicações que podem ser descritas através deste modelo, a maioria dos autores apenas se refere ao estudo do modelo Poisson-Gama. Um caso especial é verificado quando o valor de p se encontra entre o intervalo $]1,2[$, permitindo aproximar-se entre a Poisson e a Gama, o que abrange um maior conjunto de distribuições, permitindo assim obter a que melhor se adapta aos dados disponíveis, ver por exemplo Frees et al. (2016a).

Suponha-se que N representa o número de sinistros de uma carteira em que N segue uma distribuição de Poisson de valor esperado λ. Seja Y_i com $i = 1, \cdots, n$ uma sequência de variáveis aleatórias i.i.d, em que cada Y_i segue uma distribuição Gama com parâmetros α e β, representando o custo de um sinistro. Então, $S = \sum_{i=1}^{N} Y_i$ é um processo de Poisson Composto, podendo também ser designado como Poisson-Gama, ver, por exemplo, Centeno (2002).

Com o objetivo de ilustrar a componente mista da distribuição Tweedie, primeiro note-se que a probabilidade do montante de indemnizações agregadas ser nulo é igual à probabilidade de não ocorrer sinistros, ou seja,

$$P[S = 0] = P[N = 0] = e^{-\lambda}$$

Por conseguinte, utilizando a probabilidade condicional a função de distribuição de S pode ser obtida a partir de:

$$P[S \leq y] = e^{-\lambda} + \sum_{n=1}^{\infty} P[S \leq y | N = n] P[N = n], \quad y \geq 0$$

Tendo em consideração que, para $N = n$, se tem

$$S_n = Y_1 + \cdots + Y_n$$

ter-se-à que $S_n \sim \text{Gama}(n\alpha, \beta)$ pois corresponde à soma de n variáveis aleatórias independentes com distribuição Gama(α, β). Para $y > 0$, a função de densidade da distribuição Tweedie, para $1 < p < 2$, é dada por, ver, por exemplo Clark e Thayer (2004):

$$f_S(y) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{\beta^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha - 1} e^{-y\beta}, \quad y \geq 0,\ \alpha, \beta > 0 \tag{3.6}$$

Numa primeira análise à equação anterior, a função densidade pode aparentar não pertencer à Família Exponencial, dada pela expressão (3.2). Com o objetivo de entender esta relação, é necessário obter as expressões para a média e variância de S, ver, por exemplo, Clark e Thayer (2004).

$$\mathbb{E}[S] = \lambda \frac{\alpha}{\beta} \qquad \text{e} \qquad \mathbb{V}[S] = \lambda \frac{\alpha}{\beta}(1 + \alpha)$$

Definindo os três parâmetros λ, α e β em função de μ, ϕ e p, tem-se:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1} \qquad \text{e} \qquad \beta = \frac{1}{\phi(p-1)\mu^{p-1}}$$

Inserindo estes novos parâmetros na equação (3.6) obtém-se:

$$f(y; \mu, p, \phi) = \exp\left[\frac{y}{-(p-1)\mu^{p-1}\phi} - \frac{\mu^{2-p}}{2-p} + c(y, \phi) \right] \tag{3.7}$$

Considerando:

$$\theta = -\frac{1}{(p-1)\mu^{p-1}} \qquad a(\phi) = \phi \qquad b(\theta) = \frac{\mu^{2-p}}{2-p}$$

e

$$c(y,\phi) = \begin{cases} 1, & \text{se} \quad y = 1; \\ \dfrac{y^{n\left(\frac{2-p}{p-1}\right)-1}}{n!\Gamma\left(n\frac{2-p}{p-1}\right)[\phi(2-p)]^n[\phi(p-1)]^{n\left(\frac{2-p}{p-1}\right)}} & \text{se} \quad y > 0. \end{cases}$$

verifica-se, desta forma, que a função de densidade da distribuição Tweedie, para $1 < p < 2$, verifica a expressão (3.2), podendo concluir-se que a distribuição Tweedie pertence à Família Exponencial, quando $1 < p < 2$.

Por conseguinte, o valor esperado e variância de S são:

$$\mathbb{E}[S] = \mu \qquad \text{e} \qquad \mathbb{V}[S] = \phi\mu^p$$

Na Figura 3.1 são apresentados seis gráficos em que cada um representa a distribuição que a Família Distribuição Tweedie assume para o respetivo valor de p.

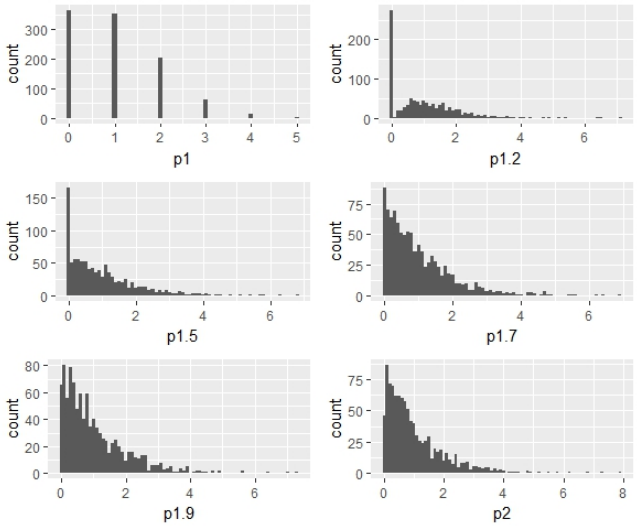

Figura 3.1: Distribuições e respetivo valor do parâmetro p

Analisando o gráfico **p1**, constata-se que a distribuição aproxima-se da distribuição de Poisson.

As distribuições presentes nos gráficos **p1.2**, **p1.5**, **p1.7** e **p1.9**, não se aproximam de distribuições conhecidas, uma vez que estas possuem a componente discreta referente à contagem dos sinistros e a parte contínua que é alusiva aos seus custos.

Quando analisado o gráfico **p2**, constata-se que a respetiva distribuição se aproxima de uma distribuição conhecida, como a distribuição Gama.

3.4 Ajustamento do modelos

3.4.1 Estimação dos parâmetros de regressão

Num modelo linear generalizado os parâmetros $\boldsymbol{\beta}$ são os parâmetros de maior interesse, os quais são estimados pelo método da máxima verosimilhança ver, por exemplo, Turkman e Silva (2000).

A função de verosimilhança do modelo, em função dos parâmetros $\boldsymbol{\beta}$, é dada por, ver por exemplo, Hosmer Jr et al. (2013) :

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\mathrm{y}_i; \theta_i, \phi) \tag{3.8}$$

Por conseguinte, a função log-verosimilhança é dada pela expressão:

$$ln\ L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \tag{3.9}$$

$$\text{com } \ell_i(\boldsymbol{\beta}) = \left[\frac{\mathrm{y}_i \theta_i - b(\theta_i)}{a(\phi)} + c(\mathrm{y}_i, \phi) \right] \text{com } i = 1, \cdots, n.$$

Uma vez que os estimadores de máxima verosimilhança para $\boldsymbol{\beta}$ são obtidos como solução do sistema de equações de verosimilhança ver, por exemplo, Cox e Hinkley (1979) , as equações de verosimilhança para $\boldsymbol{\beta}$ são dadas por:

$$\sum_{i=1}^{n} \frac{(\mathrm{y}_i - \mu_i)}{\mathbb{V}(Y_i)} \frac{\partial \mu_i}{\partial h(\mu_i)} x_{ij} = 0\,, \qquad j = 1, \cdots, p. \tag{3.10}$$

3.4.2 Testes de Hipóteses

Os testes estatísticos mais conhecidos para avaliar o nível de significância dos parâmetros de regressão (β_i ; $\quad i = 1, \cdots, n$), ver Turkman e Silva (2000), são o Teste de *Wald*, o Teste de *Wilks* e o Teste de *Rao*[1] sendo, porém, os dois primeiros os mais utilizados.

Os testes de hipóteses aqui apresentados estabelecem como hipóteses:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad vs. \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b} \tag{3.11}$$

onde $\mathbf{C}$ corresponde a uma matriz $q \times p$, $q \leq p$ e $\mathbf{b}$ é um vetor conhecido de dimensão q.

Em particular, quando o teste pretende avaliar a nulidade de um dado parâmetro β_j, têm-se as hipóteses:

[1] O Teste de Rao pode ser encontrado com detalhe em Rao (2009).

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0\,, \qquad j = 1, \cdots, p \tag{3.12}$$

A título de exemplo de aplicação dos testes de hipóteses, suponha-se que para a modelação do número de sinistros, se pretende decidir que covariáveis devem constar no modelo. Por outras palavras, pretende-se testar a hipótese se a variável é estatísticamente significativa para o modelo.

3.4.2.1 Estatística de *Wald*

O teste de *Wald* é mais utilizado para testar hipóteses nulas sobre componentes individuais baseando-se na normalidade assimptótica do estimador de máxima verosimilhança, ver Turkman e Silva (2000).

Sendo $\widehat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ e $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ a matriz de covariâncias, tem-se que:

$$\mathbf{C}\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N_p\Big(\mathbf{C}\boldsymbol{\beta}, \mathbf{C}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}^T\Big) \tag{3.13}$$

Assim, a estatística de *Wald*, sob H_0, é dada por:

$$W = \Big(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{b}\Big)^T \Big[\mathbf{C}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{C}^T\Big]^{-1}\Big(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{b}\Big) \quad \overset{a}{\sim} \chi^2_q \tag{3.14}$$

Por conseguinte, rejeita-se a hipótese nula, com um nível de significância α, se o valor observado da estatística de *Wald* for superior ao quantil de probabilidade $1-\alpha$ da distribuição χ^2_q.

3.4.2.2 Estatística de *Wilks*

O Teste de *Wilks*, ou Teste de Razão de Verosimilhanças, ver Wilks (1938), é definida através da expressão:

$$\Lambda = -2ln\left(\frac{M_0 L(\boldsymbol{\beta})}{M_1 L(\boldsymbol{\beta})}\right) \tag{3.15}$$

em que M_0 representa o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança em H_0 e M_1 o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança em H_1. Como $H_0 \subset H_1$ então $M_0 L(\boldsymbol{\beta}) \leq M_1 L(\boldsymbol{\beta})$, e por conseguinte $0 \leq \Lambda \leq 1$. Do desenvolvimento da expressão (3.15), resulta que

$$\Lambda = -2\Big(\ell(\boldsymbol{\beta}) - \ell(\widehat{\boldsymbol{\beta}})\Big)$$

Cox e Hinkley (1979) afirmam que, no Teorema de *Wilks*, a estatística Λ sob H_0, segue uma distribuição assimptótica de um χ^2, sendo o número de graus de liberdade igual à diferença entre o número de parâmetros a estimar sob H_1 (p) e o número de parâmetros a estimar sob H_0 $(p-q)$.

Assim,

$$\Lambda = -2\Big(\ell(\boldsymbol{\beta}) - \ell(\widehat{\boldsymbol{\beta}})\Big) \overset{a}{\sim} \chi^2_q \tag{3.16}$$

Desta forma, a hipótese nula é rejeitada a um nível de significância α, se o valor da estatística Λ for superior ao quantil de probabilidade 1-α da distribuição χ^2_q.

3.4.3 Intervalos de Confiança

Para obter intervalos de confiança para os parâmetros do vetor $\boldsymbol{\beta}$, existe a necessidade de conhecer a distribuição por amostragem de $\boldsymbol{\beta}$. Segundo Turkman e Silva (2000), não é possível obter as distribuições por amostragem exatas para os estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ pelo que, usualmente, se recorrem a resultados assimptóticos, que se garantem quando os modelos em estudo satisfazem as condições usuais de regularidade, o que acontece para os modelos lineares generalizados, ver por exemplo Fahrmeir e Kaufmann (1985). Assim sendo, sabendo que os estimadores são assimptóticamente centrados, ou seja:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] \approx \boldsymbol{\beta}$$

tem-se que:

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N_p\Big(\beta, \mathbf{I}^{-1}(\boldsymbol{\beta})\Big) \tag{3.17}$$

Tendo em consideração as equações (3.13) e (3.17), é possível construir intervalos de confiança para os parâmetros estimados.

No caso particular de uma única componente do vetor $\boldsymbol{\beta}$, tem-se:

$$\widehat{\beta_j} \overset{a}{\sim} N\Big(\beta_j, v_{jj}\Big) \tag{3.18}$$

em que v_{jj} representa o elemento (j, j) da matriz $\mathbf{I}^{-1}(\boldsymbol{\beta})$.

Por conseguinte, um intervalo de confiança de nível $(1-\alpha)$ para $\widehat{\beta_j}$, $j = 1, \cdots, p$ é dado por:

$$\mathbf{IC}_{1-\alpha}(\beta_j) = \widehat{\beta_j} \pm z_{\alpha/2}\sqrt{v_{jj}} \tag{3.19}$$

em que $z_{\alpha/2}$ corresponde ao quantil de probabilidade $(1 - \alpha/2)$ da distribuição Normal Reduzida.

A construção de intervalos de confiança para os parâmetros $\boldsymbol{\beta}$ será uma mais valia para avaliar a precisão das estimativas anteriormente efetuadas.

Do ponto de vista de uma seguradora, os intervalos de confiança servem para avaliar a precisão das estimativas obtidas para a construção tarifa.

3.4.4 Estimação do parâmetro de dispersão

Em certas distribuições, o parâmetro de dispersão é conhecido (por exemplo, na distribuição de Poisson), porém, de um modo geral, este parâmetro é desconhecido *a priori*, sendo assim necessário estimar o parâmetro de dispersão ϕ.

Existem diferentes soluções para a estimação do parâmetro de dispersão, porém, Turkman e Silva (2000) afirmam que o método mais simples baseia-se na estatística de *Pearson* generalizada

$$\sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{\mathbb{V}(\widehat{\mu_i})}$$

Mostra-se, ver, por exemplo, Turkman e Silva (2000), que um possível estimador para ϕ é dado por:

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \omega_i \frac{(y_i - \widehat{\mu_i})^2}{\mathbb{V}(\widehat{\mu_i})} \tag{3.20}$$

tendo-se que

- $\widehat{\phi}$ é um estimador consistente de ϕ;
- $\widehat{\phi}$ é um estimador assimptoticamente centrado;
- $\frac{\chi^2}{\phi} \overset{a}{\sim} \chi^2_{n-p}$

3.5 Seleção dos Modelos

Em modelação estatística, e em particular, na construção da tarifa, selecionar o modelo que melhor descreva os dados, é uma fase importante desta área. Quando se está perante inúmeras variáveis que podem ser importantes para descrever a variável resposta Y, o processo de seleção pode ser moroso. Uma das metodologias possíveis para escolha das variáveis significativas a reter no modelo é o método *stepwise*.

3.5.1 Forward Stepwise

Neste método, o processo é iniciado através do modelo nulo - modelo em que se considera apenas um parâmetro, que representa a média μ, de todas as observações y_i. A seleção das variáveis que melhor explicam a variável dependente Y é efetuada sequencialmente. Em cada passo, a variável mais significativa, ou seja, a variável que, através do teste de *Wald* apresente um menor *p-value*, é incluída no modelo em análise.

Após a escolha da variável a entrar no modelo, faz-se análise ao seu grau de significância através do valor do *p-value* do teste de verosimilhança entre os modelos que a incluem e a excluem, e assim se toma uma decisão acerca da exclusão (ou inclusão) da variável em estudo no modelo final ver, por exemplo, Guerreiro (2016).

3.5.2 Backward Stepwise

No método *Backward Stepwise*, as variáveis são retiradas de forma sequencial do modelo completo - o modelo em que o número de observações iguala o número de parâmetros, ou seja, o modelo em que o valor médio μ_i coincide com o valor observado y_i, e de acordo com os resultados do teste de *Wald*, retira-se a que apresentar maior *p-value*, sendo essa a menos significativa.

Após isso, através do teste de *Wilks*, é comparado o ajuste do modelo completo, isto é com o obtido após a exclusão da variável. Se o valor do *p-value* resultante do teste anterior, for inferior a um determinado nível de significância α, considera-se o modelo com a variável em estudo como sendo o melhor modelo, caso contrário a variável é retirada.

Quando se obtém um modelo em que todas as variáveis sejam significativas segundo o teste de *Wald*, o processo termina e o modelo final é constituído por todas as variáveis dessa etapa.

3.5.3 Método AIC

O método Akaike Information Criterion (AIC), foi proposto por Akaike (1974). A ideia por detrás deste critério é examinar a complexidade do modelo e avaliar como se ajusta aos dados em estudo, ver por exemplo Kaas et al. (2008).

Baseando-se na função log-verosimilhança, juntamente com a introdução de um fator de correção associada ao número de parâmetros do modelo, a medida AIC é dado por:

$$AIC = -2\ell(\boldsymbol{\beta}) - 2p \tag{3.21}$$

em que p representa o número de parâmetros do modelo e $\ell(\beta)$ o valor da função log-verosimilhança do modelo.

Considera-se como indicativo de um melhor ajustamento um valor baixo de AIC, tendo-se como objetivo na seleção de modelos a minimização deste valor.

É importante realçar que este critério é um teste entre modelos, ou seja, apenas serve para indicar qual o melhor modelo entre um conjunto de modelos dados.

> *"Of course, models not in the set remain out of consideration. AIC is useful in selecting the best model in the set; however, if all the models are very poor, AIC will still select the one estimated to be best, but even that relatively best model might be poor in an absolute sense. Thus, every effort must be made to ensure that the set of models is well founded."*
>
> — Burnham e Anderson (2003)

3.5.4 Método BIC

O método de seleção Bayesian Information Criterion (BIC), foi proposto por Schwarz, ver Schwarz (1978), e é definido como:

$$BIC = -2\ell(\boldsymbol{\beta}) - p \cdot \log(n) \tag{3.22}$$

em que $\ell(\boldsymbol{\beta})$ é o valor da função log-verosimilhança do modelo, n o número de observações e p o número de parâmetros do modelo.

Tal como no método AIC, tem-se como objetivo na seleção de modelos a minimização do valor de BIC, considerando-se apenas como indicativo de um melhor ajustamento.

3.6 Qualidade dos modelos

Após se selecionar o modelo que melhor se ajuste aos dados, é necessário aferir a qualidade desse mesmo ajustamento. Com essa finalidade, usam-se principalmente os desvios e noções relativas a resíduos do modelo.

3.6.1 A *Deviance*

A *deviance* (ou função desvio) tem como objetivo determinar a distância, ou seja, a diferença, entre os valores ajustados pelo modelo e os valores observados. Por outras palavras,compara-se o modelo em estudo (ou modelo estimado) com o modelo completo (ou modelo saturado).

Considere-se $\widehat{\boldsymbol{\beta}_C}$ e $\widehat{\boldsymbol{\beta}_E}$ as estimativas de máxima verosimilhança para o modelo completo e o modelo em estudo respetivamente.

O desvio é, portanto, segundo Fahrmeir e Tutz (2001), como uma medida de afastamento entre o modelo completo e o modelo em estudo, calculado como:

$$D^* = -2\Big(\ell(\widehat{\boldsymbol{\beta}_E}) - \ell(\widehat{\boldsymbol{\beta}_C})\Big) = -2\sum_{i=1}^{n}\Big(\ell_i(\widehat{\mu_i}) - \ell_i(\mathrm{y}_i)\Big) \tag{3.23}$$

Muitos autores consideram a substituição $a(\phi) = \dfrac{\phi}{\omega}$ em que ω_i é uma constante. No âmbito da tarifação , ω_i pode representar a exposição ao risco de cada apólice i com $i = 1, \cdots, n$. Consequentemente, tem-se:

$$D^* = -2\sum_{i=1}^{n}\frac{\omega_i}{\phi}\Big[\mathrm{y}_i\big(\widehat{\mu_i} - \mathrm{y}_i\big) - \Big(b(\theta_{i_E}) - b(\theta_{i_C})\Big)\Big] = \frac{1}{\phi}D \tag{3.24}$$

sendo D dado por:

$$D = 2\sum_{i=1}^{n}\omega_i\Big[-\mathrm{y}_i\big(\widehat{\mu_i} - \mathrm{y}_i\big) - \Big(b(\theta_{i_C}) - b(\theta_{i_E})\Big)\Big] \tag{3.25}$$

As expressões D^* e D definidas anteriormente em (3.23) e (3.25), representam o desvio reduzido e o desvio, respetivamente.

Assim, para avaliar a qualidade do ajustamento dos modelos, pode ser utilizada o valor da *deviance* reduzida, tendo em atenção que, ver por exemplo, Turkman e Silva (2000):

$$D^* \overset{a}{\sim} \chi^2_{n-p} \tag{3.26}$$

Para um nível de significância α, rejeita-se a hipótese de que o modelo em estudo é um bom ajustamento ao modelo completo se o valor de D^* for superior ao quantil de probabilidade 1-α de um χ^2_{n-p} ver, por exemplo, De Jong e Heller (2008).

Muitos autores recomendam prudência quanto ao uso da *deviance*, ver por exemplo McCullagh (1984), uma vez que para aferir a qualidade do ajustamento depende de pressupostos que não são frequentemente sustentáveis.

Como já referido, o parâmetro de dispersão ϕ nem sempre é conhecido, e como tal necessita de ser estimado. Uma vez estimado, a aproximação de D^* à distribuição χ^2_{n-p} pode ficar comprometida, ver, por exemplo, De Jong e Heller (2008).

No caso da distribuição de Poisson, o parâmetro de dispersão é conhecido, tomando o valor 1 e a aproximação à distribuição χ^2 é bastante útil. Por outro lado, na distribuição Normal, quando σ^2 é conhecido a distribuição de χ^2 é exata, porém, quando o σ^2 necessita de ser estimado, a aproximação é pouco fiável.

3.6.2 Resíduos

A análise de resíduos, ver, por exemplo, Turkman e Silva (2000), é útil, não só para uma avaliação local da qualidade de ajustamento de um modelo no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear, como também para ajudar a identificar observações mal ajustadas, i.e., que não são bem explicadas pelo modelo.

Um resíduo R_i deve exprimir a discrepância entre o valor observado y_i e o valor $\widehat{y}_i$ ajustado pelo modelo ver, por exemplo, Turkman e Silva (2000) .

3.6.2.1 Resíduos de *Pearson*

A escolha mais comum para avaliação dos resíduos das observações corresponde aos Resíduos de *Pearson*, definidos por:

$$R_i^p = \frac{y_i - \widehat{\mu_i}}{\sqrt{\mathbb{V}(\widehat{\mu_i})}} \tag{3.27}$$

O resíduo R_i^p, assim definido, corresponde à contribuição de cada observação para o cálculo da estatística de *Pearson* generalizada.

3.6.2.2 Resíduos do Desvio

Outros resíduos utilizados são os resíduos da *deviance*, onde é utilizada a i-ésima observação da função desvio, dada pela equação (3.23). Assim, de acordo com Wood (2006), o

Desvio Residual é dado por:

$$R_i^D = \text{sinal}(y_i - \widehat{\mu})\sqrt{d_i} \tag{3.28}$$

onde d_i corresponde ao desvio de cada observação, ou seja, corresponde à diferença entre os logaritmos das verosimilhanças observada e ajustada para cada observação.

Os resíduos são regra geral, observados graficamente, analisando o gráfico dos valores dos resíduos contra os valores ajustados, o que nos permite também analisar mais facilmente a existência de *outliers* que poderão não ser incluídos no modelo.

4

Construção de uma Tarifa de Responsabilidade Civil Automóvel

A componente prática desta dissertação começa por uma breve descrição da empresa Actuariado, a qual facultou os dados para a realização desta dissertação, dos seus serviços e áreas envolventes, bem como os objetivos a que esta tese se propõe. De seguida, é realizado um tratamento estatístico aos dados da carteira analisada, de forma a serem coerentes e consistentes para o estudo a realizar. Posteriormente, são aplicadas as metodologias, tal como descritas no Capítulo 3, de forma a selecionar o modelo que melhor se ajuste aos dados da sinistralidade da carteira, o que permitirá a obtenção da tarifa de responsabilidade civil automóvel. Por último, é feita uma análise aos resultados obtidos.

Descrição da Empresa e Objetivos

A Actuariado (www.actuariado.pt), empresa onde foi realizado o estágio curricular, dedica-se a estudos atuariais relacionadas com a atividade seguradora e com a avaliação de passivos de empresas. Foi fundada em 1988 e as suas principais áreas de negócios são:

• Consultoria Atuarial - Realização de avaliações actuariais de provisões técnicas de carteiras de seguradoras do ramo vida e não vida; em validações da *best estimate liability* (BEL), do *solvency capital requirement* (SCR), de acordo com a legislação de Solvência II, e em trabalhos de *Pricing*.

• Avaliação de Seguradoras - Avaliação económica e de passivos de empresas seguradoras.

• Fundos de Pensões – Desenho e avaliação das responsabilidades em Planos e Fundos de Pensões; controlo da evolução das responsabilidades futuras das empresas. Outras das competências técnicas que a Actuariado dispõe são os processos de constituição de seguradoras, auditoria, bem como formação atuarial.

Os dados facultados pela Actuariado correspondem a uma carteira automóvel de uma Seguradora do ramo não vida, a operar em Portugal. Por motivos de confidencialidade, apenas se pode referir esta dispõe de diferentes linhas de negócio, apesar do presente estudo incidir apenas na área de seguros automóveis. Um dos objetivos da Actuariado é desenvolver e aprofundar conhecimentos técnicos para que no futuro possam ser utilizadas novas metodologias na área da tarifação automóvel. Esta área envolve várias coberturas como por exemplo, a cobertura por Danos Próprios, Responsabilidade Civil, Quebra Isolada de Vidros, Furto, entre outras. Neste caso em particular, esta dissertação apenas se foca na cobertura de Responsabilidade Civil.

4.1 Descrição dos dados

Numa fase inicial é importante realizar uma análise detalhada dos dados da carteira em estudo.

Após uma primeira análise à base de dados fornecida, foi realizado um tratamento dos dados de modo a retirar dados inconsistentes bem como repetições que pudessem provocar um enviesamento dos resultados.

Todo este trabalho inicial é essencial para se obter uma base de dados final com informação considerada necessária e relevante para análise em estudo. A base de dados contém informação privada de segurados bem como os respetivos dados de apólice e sinistros.

Na base de dados considerada como final para o estudo existem no total 125.857 linhas (registos) e 24 colunas (variáveis), porém apenas 14 destas serão utilizadas para o estudo (ver Tabela 4.1). As variáveis podem ser agrupadas em 5 grupos: variáveis de controlo, características dos segurados, variáveis geográficas, características do veículo e variáveis de resposta.

De referir que uma apólice na base de dados pode conter mais de um sinistro. A variável *ClaimCount* representa o número total de sinistros, a variável *TotalCost* faz referência ao custo total de cada sinistro registado e provém da soma das variáveis *ClaimCost_TPL_MD* e *ClaimCost_TPL_BI*, que representam o custo da cobertura de responsabilidade civil para dados materiais e corporais, respetivamente.

Tabela 4.1: Variáveis de estudo

Controlo	Segurado	Veículo	Geográfica	Resposta
Exposição	IdadeSegurado	IdadeVeículo	Localidade	ClaimCount
Num.id	AnosCarta	Marca		TotalCost
		TipoVeiculo		
		PesoVeículo		
		NLugares		
		CapCubVeiculo		
		Combustível		
		TipoUtilização		

As variáveis *Exposição* e *Num.id* representam o tempo (fração do ano) em que o veículo esteve exposto ao risco e o número da apólice, respetivamente.

Na base de dados está presente apenas uma variável de cariz geográfico, *Localidade*, que indica a cidade/distrito em que o segurado reside.

As variáveis acerca do segurado correspondem às suas principais características, a Idade (*IdadeSegurado*) e o Tempo de Carta de Condução (*AnosCarta*), medida em anos.

As características do veículo correspondem a atributos como a Idade (*IdadeVeículo*), o Número de Lugares (*NLugares*), o Peso do Veículo (*PesoVeículo*). As variáveis *TipoVeículo* e *TipoUtilização*, correspondem , como a designação indica, ao Tipo de Veículo (carro, motociclo ou camião) e ao Tipo de Utilização (pessoal ou privado), respetivamente. Existem também variáveis relacionadas com o motor, como a Capacidade Cúbica (*CapCubVeiculo*) e o Combustível *Combustível* (Diesel, Gasolina, Gás ou Elétrico).

Aquando da modelação da severidade dos sinistros, a variável resposta é a variável contínua *TotalCost*, enquanto que, para a modelação da frequência de sinistralidade, a variável resposta é a variável discreta *ClaimCount*.

É importante efetuar uma prévia às variáveis *ClaimCount* e *TotalCost*. Com estas variáveis, é possível analisar algumas percentagens úteis para criar um modelo tarifário.

Tabela 4.2: Dados da carteira

Apólices	**Total Sinistros**	**Custo Total (€)**
74.695	4.860	6.970.070,04

Tabela 4.3: Estatística base

Frequência de Sinistralidade	**Custo Médio por apólice**	**Desvio Padrão do custo por apólice**	**Custo Médio por sinistros**	**Desvio Padrão do custo por sinistro**
8,87%	93,31€	1.889,25€	1.434,17€	7.261,84€

Analisando a Tabela 4.2, é possível verificar que na carteira em estudo foram declarados um total de 4.860 sinistros e que estes representam um custo total de cerca de 6.970.070 euros.

Relativamente à Tabela 4.3, verifica-se que a carteira apresenta uma frequência de sinistralidade em torno dos 8,9%.

Tendo em conta o número elevado de apólices na carteira e o número de sinistros declarados, constata-se que o custo médio por apólice de 93,31 euros. Por outro lado, o custo médio de um sinistro declarado é de 1.434,17 euros.

Uma vez que na carteira em estudo foram observados 4.860 sinistros, é importante analisar como estão distribuídos por apólices e como os custos associados a estes sinistros se distribuem ao nível dos quantis.

Tabela 4.4: Distribuição de sinistros por apólices

Sinistros	0	1	2	3	4	5
Apólices	70.140	4.272	265	15	2	1

Tabela 4.5: Quantis de custos totais (€)

50 %	90%	95%	99%
773,90	2.530,71	4.016,37	11.691,74

Realizando uma análise às Tabelas 4.4 e 4.5 conclui-se que apenas em 4.555 das 74.695 apólices foram declarados sinistros e que 99% destes sinistros custaram no máximo 11.691,74 euros.

4.1.1 Análise às variáveis tarifárias

Para realizar uma análise preliminar à carteira em estudo, é fulcral analisar individualmente cada variável, visto ser um processo essencial de modo a possibilitar uma modelização mais precisa quer da severidade quer da frequência de sinistralidade.

Idade do Segurado

A variável *IdadeSegurado*, apresenta valores entre os 17 e os 110 anos. Uma vez que não se encontra dividida em classes etárias, e de modo a criar um modelo de estrutura tarifária com esta variável, é necessário realizar uma partição em classes.

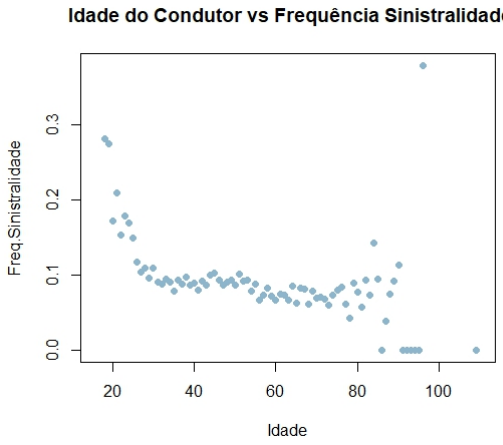

Figura 4.1: Frequência de sinistralidade empírica vs Idade do Segurado.

Na Figura 4.1, são evidentes as diferenças na frequência de sinistralidade em função de idade do segurado. As diferenças significativas entre segurados mais jovens e menos jovens reflete a necessidade de tarifação distinta de acordo com a classe etária do segurado. Consequentemente, definem-se 4 escalões para a variável em estudo:

Tabela 4.6: Níveis tarifários para a variável *IdadeSegurado*

Nível Tarifário	**Escalões de idades**
IDA1	]16, 30]
IDA2	]30, 55]
IDA3	]55, 75]
IDA4	]75, 110]

Após a definição dos diferentes escalões tarifários, é relevante avaliar graficamente a frequência de sinistralidade empírica nos diferentes níveis tarifários.

O gráfico da Figura 4.2 ilustra ainda um intervalo de confiança para a frequência de sinistralidade estimada em cada um dos escalões utilizados.

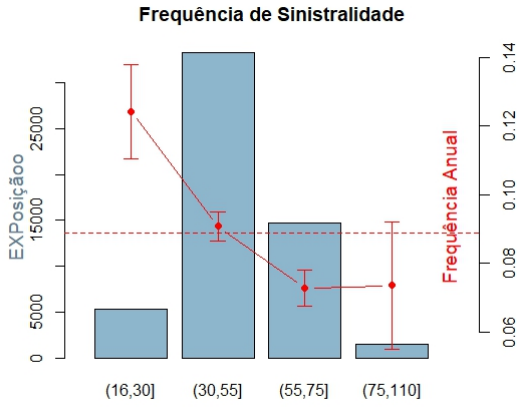

Figura 4.2: Frequência de sinistralidade empírica por nível tarifário Idade do Segurado.

Ao analisar-se o gráfico da Figura 4.2, verifica-se que existem diferenças significativas na frequência de sinistralidade nos três primeiros níveis tarifários.

A título de exemplo, a frequência de sinistralidade na faixa etária dos 17 aos 30 anos (0,124) é mais elevada do que a apresentada pelos segurados da faixa etária dos 31 aos 55 anos (0,091), sendo esta última muito próxima da média da carteira.

Os segurados com idades acima dos 55 anos apresentam frequência de sinistralidade média abaixo da média da carteira, devendo portanto ser tarifados com um prémio mais baixo.

Faz-se notar, desde já, que a frequência de sinistralidade observada para os segurados do escalão]75,110] é muito idêntica à do escalão]55,75]. Adicionalmente, o mínimo de apólices existente neste escalão é muito reduzido, o que poderá comprometer a precisão das estimativas a efetuar neste intervalo, conforme se pode desde já observar pela amplitude do intervalo de confiança.

Localidade

Na variável tarifária *Localidade*, são considerados os 18 distritos de Portugal continental mais as ilhas dos Açores e da Madeira.

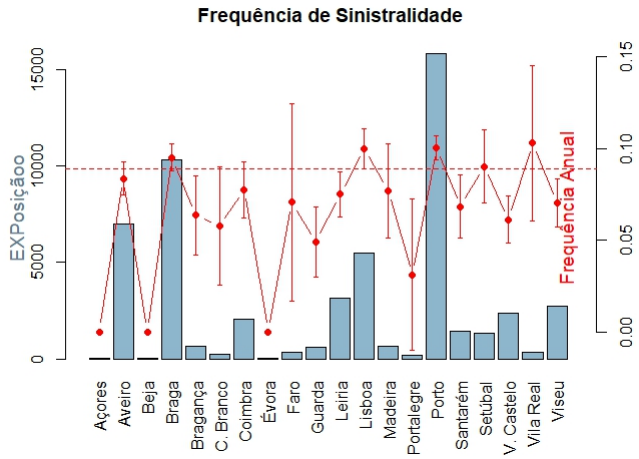

Figura 4.3: Frequência de sinistralidade empírica por Localidde.

Com base no gráfico da Figura 4.3, pode-se analisar que alguns distritos apresentam frequências de sinistralidade semelhantes e que alguns dos distritos contêm um número muito reduzido de apólices, o que não é desejável de um ponto de vista estatístico. Assim, definem-se 6 escalões tarifários:

Tabela 4.7: Níveis tarifários para a variável *Localidade*

Nível Tarifário	**Distritos**
Loc1	Lisboa, Setúbal e Vila Real
Loc2	Coimbra, Leiria, Madeira, Santarém, Viana do Castelo e Viseu
Loc3	Açores, Beja, Bragança, Castelo Branco, Évora, Faro, Guarda e Portalegre
Loc4	Porto
Loc5	Braga
Loc6	Aveiro

Idade do Veículo

A variável *IdadeVeículo*, apresenta valores entre os 0 e 58 anos. A frequência de sinistralidade empírica para esta variável é apresentada na Figura 4.4.

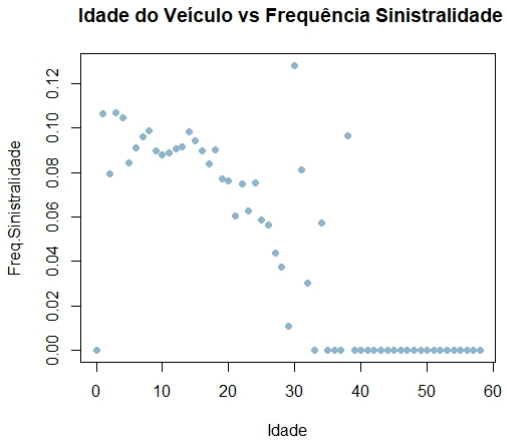

Figura 4.4: Frequência de sinistralidade empírica vs Idade do Veículo

Com base na análise o do gráfico da Figura 4.4 constata-se efetivamente as diferenças da frequência de sinistralidade em função da idade do veículo. As diferenças entre veículos novos e veículos mais velhos, podendo ser indicativo de que os carros com mais anos vão sendo cada vez mais raros e/ou cada vez menos utilizados no dia a dia. Por conseguinte, constrói-se escalões tarifários de modo a criar grupos homogéneos de risco. A opção efetuada relativamente aos escalões para esta variável são apresentados na Tabela 4.8.

Tabela 4.8: Níveis tarifários para variável *IdadeVeiculo*

Nível Tarifário	**Escalões de idades**
IDAV1	$[0,5]$
IDAV2	$]5,10]$
IDAV3	$]10,15]$
IDAV4	$]15,30]$
IDAV5	$]30,58]$

Utilizando a mesma metodologia utilizada na variável *IdadeSegurado*, é necessário construir um gráfico que expresse a frequência de sinistralidade por escalão.

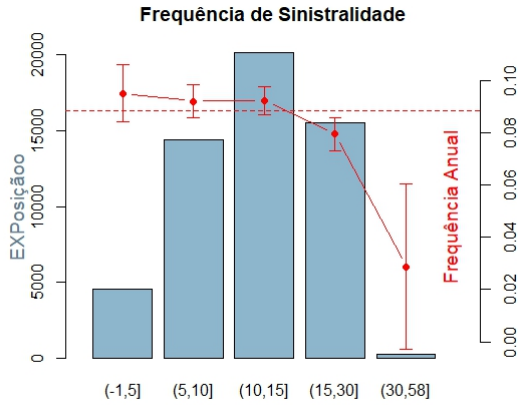

Figura 4.5: Frequência de sinistralidade empírica por nível tarifário Idade do Veículo

Analisando o gráfico da Figura 4.5 constata-se que três dos cinco escalões apresentam frequências de sinistralidade semelhantes entre si e superiores à média de sinistralidade da carteira. Em contrapartida, o último escalão apresenta uma baixa frequência de sinistralidade, podendo ser uma consequência da existência de poucas apólices para estudo, neste escalão tarifário.

Anos de Carta de Condução

A variável *AnosCarta*, apresenta idades até aos 71 anos e, à semelhança das outras variáveis numéricas apresentadas, necessita ser agrupada em escalões tarifários.

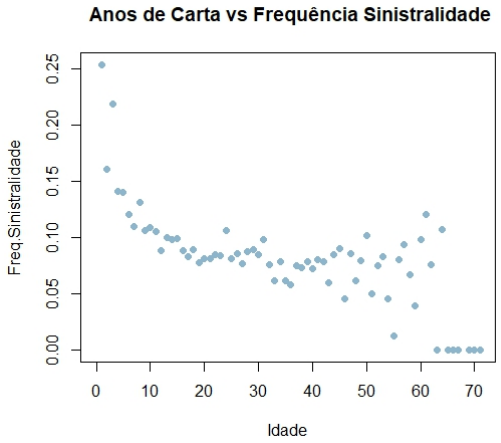

Figura 4.6: Frequência de sinistralidade empírica vs Anos de Carta.

Analisando a frequência de sinistralidade presente na Figura 4.6, faz-se notar as diferenças existentes na frequência de sinistralidade consoante os anos de carta de condução. São evidentes a diferenças entre segurados com poucos anos de condução e segurados

com mais anos de experiência, tornando-se necessário efetuar tarifações distintas. Consequentemente definem-se 4 escalões para a modelação da presente variável:

Tabela 4.9: Níveis tarifários para a variável *AnosCarta*

Nível Tarifário	**Escalões de idades**
AC1	]0,15]
AC2	]15,30]
AC3	]30,50]
AC4	]50,71]

Com base na Tabela 4.9 é apresentada a frequência de sinistralidade presente em cada escalão tarifário na Figura 4.7.

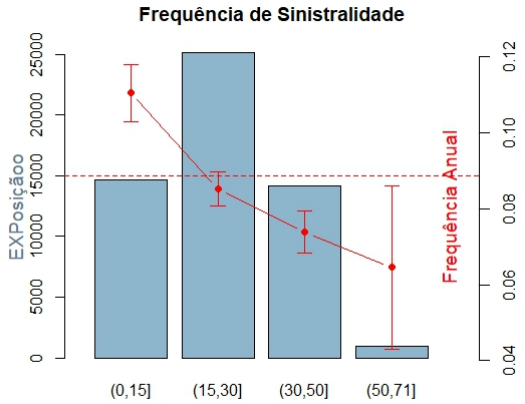

Figura 4.7: Frequência de sinistralidade empírica por nível tarifário Anos de Carta.

Analisando o gráfico da Figura 4.7 verifica-se que o escalão tarifário (0,15] (AC1) apresenta uma maior frequência de sinistralidade comparativamente aos restantes escalões, sendo o único a apresentar uma sinistralidade superior à sinistralidade média da carteira. Esta observação é consistente com o observado na Figura 4.2, sendo que usualmente são os segurados mais jovens que possuem carta de condução à menos tempo, podendo significar inexperiência e, portanto a magnitude da frequência de sinistralidade no escalão AC1 era expectável.

Tipo de Utilização

A variável *TipoUtilização* é dividida em 2 grupos:

Tabela 4.10: Níveis tarifários para a variável *TipoUtilização*

Nível Tarifário	Tipo de Cliente
T1	*Company*
T2	*Personal*

Como se pode verificar na tabela 4.10, a divisão em 2 escalões deve-se ao tipo de proprietário; *Company*, se a viatura pertence a uma empresa, ou seja, a utilização é com o intuito profissional, e *Personal* se a viatura é particular, o que indica um uso a nível pessoal. De seguida, na Figura 4.8 apresenta-se o gráfico da frequência de sinistralidade empírica por tipo de utilização.

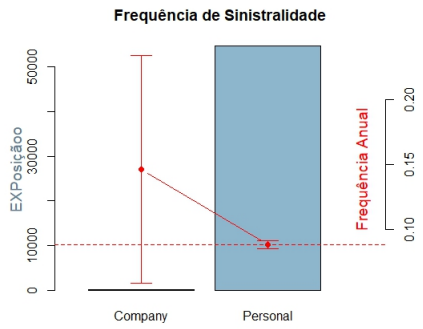

Figura 4.8: Frequência de sinistralidade empírica por Tipo de Utilização.

Analisando o gráfico da Figura 4.8, os veículos de carácter pessoal apresentam uma frequência de sinistralidade ligeiramente superior à frequência média da carteira, mas muito inferior à frequência de sinistralidade dos veículos de empresas.

Adicionalmente, faz-se notar que o número de uso empresarial é muito reduzido na carteira em análise.

Combustível

No presente estudo, existem 4 escalões para o tipo de combustível utilizado pelas viaturas, conforme se pode verificar na Figura 4.9.

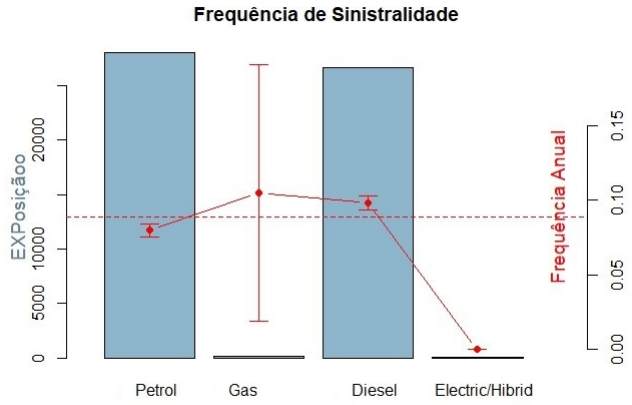

Figura 4.9: Frequência de sinistralidade empírica por Combustível.

Analisando o gráfico da Figura 4.9, o grupo das viaturas elétricas e as viaturas a gás, apresentam frequências de sinistralidade diferentes, contudo, como ambos apresentam muito poucas apólices, procedeu-se à junção do escalão das viaturas a gás ao escalão das viaturas a diesel, devido ao facto de ambas apresentarem uma frequência de sinistralidade superior à frequência média da carteira. Uma vez que as viaturas movíveis a eletricidade e as viaturas a gasolina apresentam frequência de sinistralidade inferior à sinistralidade média da carteira, efetuou-se a junção destes dois tipos de combustível num só escalão.

Assim, na Tabela 4.11 são apresentados os escalões tarifários para esta variável.

Tabela 4.11: Níveis tarifários para a variável *Combustível*

Nível Tarifário	**Combustível**
C1	Gasolina/Elétrico
C2	Diesel/Gás

Tipo de Veículo

Na Tabela 4.12 são apresentados os 3 níveis tarifários existentes para a presente variável.

Tabela 4.12: Níveis tarifários para a variável *TipoVeículo*

Nível Tarifário	**Grupo**
G1	Carros
G2	Motociclos
G3	Camiões

De seguida, na Figura 4.10 é apresentado o gráfico da frequência de sinistralidade por tipo de veículo bem como a respetiva exposição.

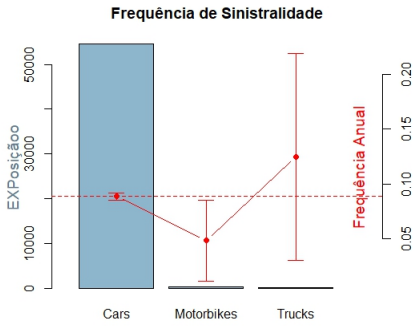

Figura 4.10: Frequência de sinistralidade empírica por Tipo de Veículo.

Através de uma análise do gráfico da Figura 4.10, verifica-se que a frequência de sinistralidade presente nos automóveis é inferior à frequência apresentada pelos veículos pesados, e superior à frequência de sinistralidade dos veículos de duas rodas.

Verifica-se ainda a existência de poucas apólices referentes a motociclos e camiões.

Número de Lugares

A variável *NLugares* representa o número de lugares das viaturas, que varia entre o mínimo de 1 e máximo de 9 lugares. Para o estudo da variável *NLugares* apresenta-se a seguinte figura:

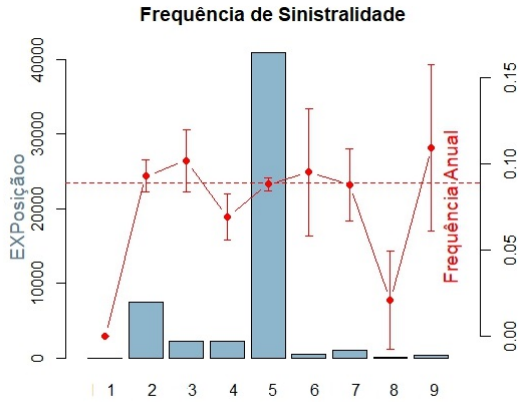

Figura 4.11: Frequência de sinistralidade empírica por Número de lugares.

Analisando o gráfico da Figura 4.11, existem vários escalões que possuem poucas apólices, como por exemplo veículos com 1 lugar, o que vai de encontro ao observado na Figura 4.10. Verifica-se que viaturas com 1, 4, 7 e 8 lugares apresentam muito poucas apólices e apresentam frequência de sinistralidade inferior à frequência média da carteira,

o que contrasta com os demais, pelo que optou-se por criar apenas 3 escalões tarifários que são apresentados na Tabela 4.13.

Tabela 4.13: Níveis tarifários para a variável *NLugares*

Nível Tarifário	Número de Lugares
L1	1, 4, 7 e 8 lugares
L2	2, 3, 6 e 9 lugares
L3	5 lugares

Capacidade Cúbica

Para a variável *CapCubVeículo*, são apresentados escalões desde o 001 que faz referência a veículos que possuem uma capacidade cúbica entre os 50 e os 125 cm^3, e ao escalão 023 que faz alusão a veículos com cilindrada entre os 5.001 e os 10.000 cm^3.

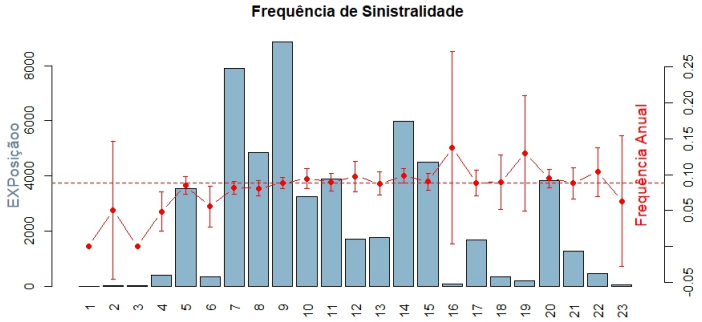

Figura 4.12: Frequência de sinistralidade empírica por Capacidade Cúbica.

Com base na Figura 4.12, verifica-se que existem escalões que apresentam frequência de sinistralidade semelhante, e por isso, optou-se por efetuar a junção de certos escalões, sendo estes definidos na Tabela 4.14.

Tabela 4.14: Níveis tarifários para a variável *CapCubVeículo*

Nível Tarifário	Escalão
CC1	010, 011, 012, 014, 015, 016, 018, 019, 020 e 022
CC2	001, 002, 003, 004, 006 e 023
CC3	005, 007, 008, 009, 013, 017 e 021

Peso do Veículo

Na presente variável, são apresentados 27 escalões de peso. O primeiro é o escalão 001, que faz referência a veículos com peso até 205 quilogramas, enquanto que o último escalão, o 27, se refere a veículos com peso superior a 3.500 quilogramas. No gráfico da Figura 4.13 apresenta-se a frequência de sinistralidade observada nos diferentes escalões.

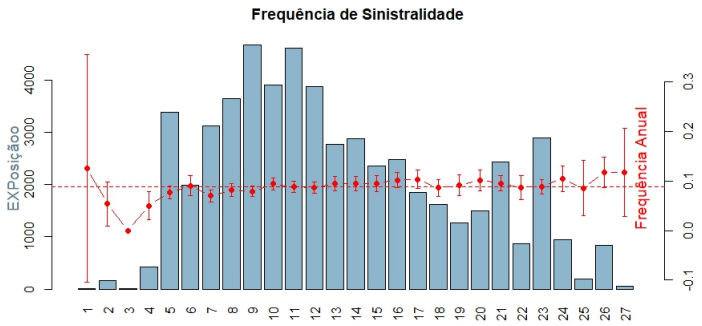

Figura 4.13: Frequência de sinistralidade empírica por Peso.

Tendo em atenção o gráfico da Figura 4.13, verifica-se que existem escalões que apresentam frequência de sinistralidade semelhante, como tal, optou-se por unir alguns desses escalões. Efetuando esse processo, consegue-se reduzir de 27 para apenas 5 escalões tarifários, sendo estes apresentados na Tabela 4.15.

Tabela 4.15: Níveis tarifários para a variável *PesoVeículo*

Nível Tarifário	**Escalão**
P1	001, 017, 019, 020, 024, 026, e 027
P2	002, 003, 004, 005, 007, 008 e 009
P3	006, 011 e 023
P4	010, 013, 014, 015, 016 e 021
P5	012, 18, 022 e 025

Na Tabela 4.16 são apresentadas todas as variáveis utilizadas na construção da tarifa automóvel. De notar que as variáveis *CapCubVeiculo*, *PesoVeiculo*, estão apresentadas categoricamente, pelo que o seu significado é apresentado em anexo. Também a variável *Marcas* é apresentado em anexo devido ao elevado número de marcas de veículos existentes na carteira em estudo.

Tabela 4.16: Variáveis Tarifárias

Fator Tarifário	Nível Tarifário	Descrição	Fator Tarifário	Nível Tarifário	Descrição
IdadeSegurado	IDA1	Dos 17 aos 30 anos	IdadeVeiculo	IDAV1	Até aos 5 anos
	IDA2	Dos 31 aos 55 anos		IDAV2	Dos 6 aos 10 anos
	IDA3	Dos 56 aos 75 anos		IDAV3	Dos 11 aos 15 anos
	IDA4	Dos 75 aos 110 anos		IDAV4	Dos 16 aos 30 anos
Localidade	Loc1	Lisboa, Setúbal e Vila Real		IDAV5	Dos 30 aos 58 anos
	Loc2	Coimbra, Leiria, Madeira, Santarém, Viana do Castelo e Viseu	AnosCarta	AC1	Até aos 15 anos
				AC2	Dos 16 aos 30 anos
	Loc3	Açores, Beja, Bragança, C. Branco, Évora, Faro, Guarda e Portalegre		AC3	Dos 31 aos 50 anos
				AC4	Dos 51 aos 71 anos
	Loc4	Porto	TipoVeiculo	G1	Carro
	Loc5	Braga		G2	Motociclos
	Loc6	Aveiro		G3	Camiões
CapCubVeiculo	CC1	010, 011, 012, 014, 015, 016, 018, 019, 020 e 022	TipoUtilização	T1	Company
				T2	Personal
	CC2	001, 002, 003, 004, 006 e 023	Combustível	C1	Diesel/Gás
	CC3	005, 007, 008, 009, 013, 017 e 021		C2	Gasolina/Elétrico
PesoVeiculo	P1	001, 017, 019, 020, 024, 026 e 027	NLugares	L1	1, 4, 7 e 8 Lugares
	P2	002, 003, 004, 005, 007, 008 e 009		L2	2, 3 6 e 9 Lugares
	P3	006, 011 e 023		L3	5 Lugares
	P4	010, 013, 014, 015, 016 e 021	Marcas	M1	Ver em anexo
	P5	012, 018, 022 e 025		M2	
				M3	
				M4	
				M5	

4.2 Modelação da Frequência de Sinistralidade

Relativamente à variável aleatória N- número de sinistros de uma apólice retirada ao acaso do conjunto de apólices da carteira, e com base nos dados anteriormente apresentados, pode verificar-se que a média e a variância da amostra são respetivamente 0,0651 e 0,0697. Dado que $\widehat{\sigma}^2 > \widehat{\mu}$, está-se perante um fenómeno de sobredispersão ver, por exemplo, De Jong e Heller (2008).

Como o objetivo de avaliar a melhor distribuição que se ajusta aos dados, a primeira a ser estudada é a distribuição de *Poisson* de parâmetro λ. Verifica-se que o valor obtido para o estimador de máxima verosimilhança para λ é igual ao valor da média observada, na amostra, pelo que $\widehat{\lambda} = \overline{n} = 0.0651$.

Na Tabela 4.17 são apresentadas as probabilidades de uma variável aleatória proveniente de uma família *Poisson* com parâmetro λ =0,0651, tal como as frequências esperadas e os valores do teste de ajustamento do Qui-Quadrado.

Tabela 4.17: Modelação do número de sinistros - Distribuição *Poisson*

i	$\mathbf{O}_i$	$\mathbf{p}_i$	$\mathbf{E}_i$	χ^2_{calc}
0	70.140	0,937	69.989,73	0,32
1	4.272	0,061	4.553,85	17,44
≥2	283	0,002	151,41	114,35
Total	74.695	1	74.695	132,12

i-N.º Sinistros O_i-N.º Apólices p_i-Probabilidade E_i-Frequência Esperada

O teste de χ^2, ver por exemplo Snedecor e Cochran (1989), é utilizado para medir a qualidade do ajustamento.

Ao analisar-se a tabela anterior, é de realçar que as classes com 2 ou mais sinistros apresentavam valores para a frequência esperada inferior ao limite mínimo exigido, neste caso 4, pelo que, a classe ≥ 2 representa as classes com 2 ou mais sinistros.

Uma vez que após o agrupamento das 4 classes acima referidas, o ajustamento obtido dos através do teste χ^2 é rejeitado a um nível de significância de 5%, sendo $\chi^2_{1;0,05} = 3,8411$, inferior ao valor obtido para χ^2_{calc}, pelo que a distribuição de *Poisson* não é a mais indicada para modelar o número de sinistros desta carteira.

A rejeição do modelo de *Poisson*, é indicativo da existência de heterogeneidade dos segurados em relação ao número de sinistros. Uma das alternativas à distribuição de Poisson, apresentadas por vários autores é a distribuição Binomial Negativa ver, por exemplo, De Jong e Heller (2008), Lemaire (1995) ou Centeno (2002).

Seguindo a mesma metodologia aquando do estudo no processo de Poisson, através do método de máxima verosimilhança, estima-se os parâmetros $\widehat{\alpha} = 0,9090$ e $\widehat{p} = 0,06506$, da Binomial Negativa, os resultados do teste estão presentes na Tabela 4.18.

Tabela 4.18: Modelação do número de sinistros - Distribuição Binomial Negativa

i	$\mathbf{O}_i$	$\mathbf{p}_i$	$\mathbf{E}_i$	χ^2_{calc}
0	70.140	0,939	70.145,46	0,00
1	4.272	0,057	4.259,17	0,04
2	265	0,003	271,55	0,13
≥ 3	18	0,000	18,82	0,06
Total	74.695	1	74.695	0,23

i-N.º Sinistros O_i-N.º Apólices p_i-Probabilidade E_i-Frequência Esperada

Uma vez que o valor observado da estatística de teste do Qui-Quadrado, χ^2_{calc}, é inferior ao valor de $\chi^2_{1;0,05} = 3,8411$, não se rejeita a hipótese de que esta distribuição se ajusta bem aos dados da carteira, ou seja, não se rejeita a hipótese de o número de sinistros ser proveniente de uma distribuição Binomial Negativa. Recorde-se que esta assumpção equivale a assumir que o número de sinistros de uma apólice retirada ao acaso do conjunto de apólices da carteira é modelado através de um processo de Poisson Misto em que a variável aleatória segue uma distribuição Gama(α, β).

Neste caso particular, tem-se como estimativas para o parâmetros de Gama, os valores $\widehat{\alpha} = 0,9090$ e $\widehat{\beta} = 14,3704$.

Nas Figuras 4.14 e 4.15 é possível verificar o ajustamento de cada distribuição aos dados de sinistralidade em estudo e corroborar os estudos estatísticos anteriormente realizados.

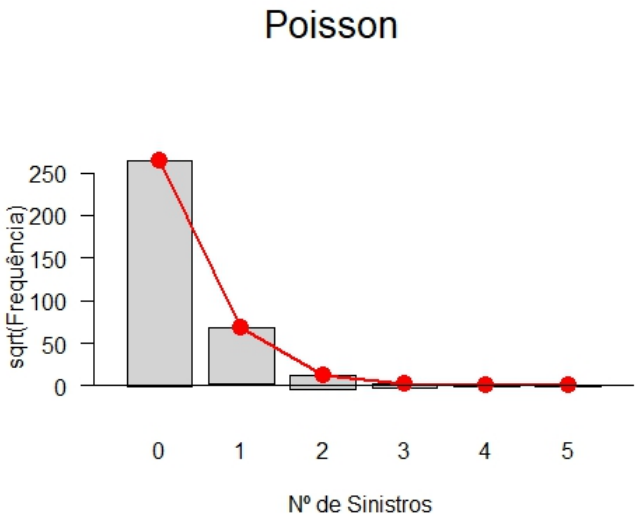

Figura 4.14: Ajustamento de uma distribuição Poisson

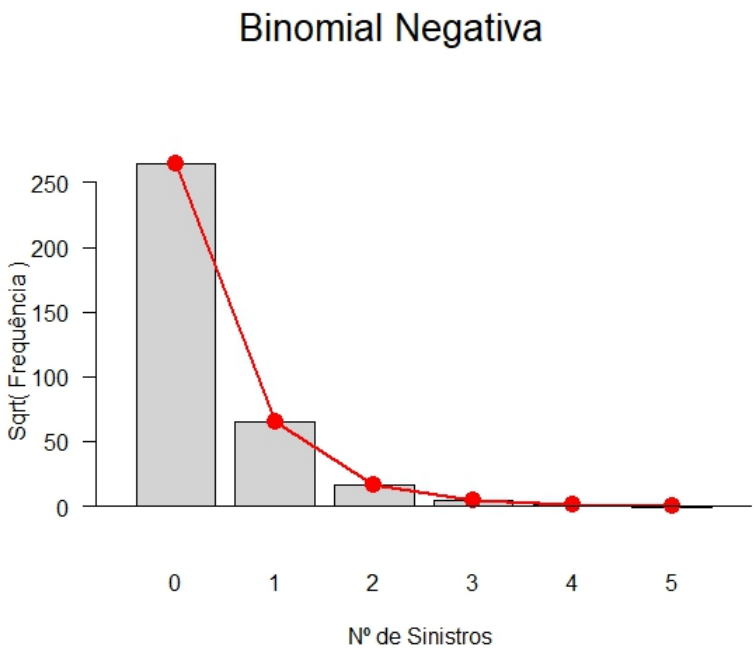

Figura 4.15: Ajustamento de uma distribuição Binomial Negativa

Concluindo que a distribuição Binomial Negativa é mais indicada para o ajustamento dos dados do número de sinistros do que a distribuição de Poisson, o próximo passo é criar um modelo que melhor descreva a frequência de sinistralidade na estrutura tarifária. Para tal, utiliza-se a função fornecida pela ferramenta estatística R:

```
glm.nb(Claims ~ Var. explicativas + offset(log(Exposicao)), data=base.dados)
```

A variável *Exposição* representa a exposição ao risco por apólice. A variável *offset*, é utilizada com o intuito de considerar que nem todos os segurados estão sujeitos ao risco durante o mesmo período temporal, e que essa informação deverá ser tida em conta no ajustamento a efetuar.

Para decidir quais as covariáveis que devem ser consideradas no modelo, a análise à nulidade dos parâmetros β pode ser realizada através de um teste de razão de verosimilhança. Assim, usa-se a função *anova* fornecida pelo R.

```
anova(modelo_n, modelo_(n-1), teste="Chisq")
```

Depois de identificadas que covariáveis devem ser consideradas, seguindo-se os procedimentos descritos na secção 3.5, é necessário analisar se todos os níveis tarifários são significativamente distintos entre si ou, em alternativa, se se pode efetuar a junção de alguns deles com o objetivo de simplificar a estrutura tarifária.

Para efetuar os testes estatísticos, recorre-se por exemplo, às funções:

```
wald.test (b=coef(modelo), Sigma=vcov(modelo), Terms=x)
```

```
linearHypothesis(modelo, "Age_Vehicle(0,2]=0")
```

Efetuando os testes acima descritos , rejeita-se qualquer cenário em teste caso os respetivos $p - value$ sejam inferiores a 5%.

No decorrer do estudo, à medida que se vão retirando variáveis que não são significativas, é necessário proceder à seleção do modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que melhor explique o comportamento da variável resposta. Como descrito no capítulo 3, um dos métodos de seleção é o critério AIC, ver secção 3.5.3. Consequentemente, e partindo do modelo inicial retirou-se as variáveis não significativas, seguindo a metodologia *Backward Stepwise*, e testou-se se os modelos obtidos apresentavam um melhor ajustamento segundo o critério anteriormente referido, obtendo desta forma um modelo final.

Na Tabela 4.19 são apresentados os valor do critério AIC para o modelo inicial e final, bem como o número de variáveis existente em cada modelo.

Tabela 4.19: Modelos frequência de sinistralidade - AIC

Modelo	N.º de variáveis	AIC
Modelo Inicial	30	35.881
Modelo Final	17	35.868

Na Tabela 4.20 é apresentado o modelo final e as estimativas dos parâmetros do modelo para cada nível tarifário, bem como o respetivo *p-value*. Este é o modelo a ser utilizado para a modelação do número de sinistros.

Tabela 4.20: Estrutura Tarifária Final - Frequência de Sinistralidade

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.3170	0.0420	-55.23	0.0000
Loc2/3	-0.3844	0.0392	-9.81	0.0000
Loc5	-0.0797	0.0397	-2.01	0.0446
Loc6	-0.1860	0.0478	-3.89	0.0001
IDAV1/2	-0.0700	0.0350	-2.00	0.0454
IDAV4/5	-0.0857	0.0381	-2.25	0.0243
IDA1	0.1629	0.0523	3.11	0.0018
IDA3/4	-0.1649	0.0367	-4.49	0.0000
AC1	0.2034	0.0391	5.20	0.0000
C1	0.1582	0.0329	4.81	0.0000
T1	0.5049	0.2830	1.78	0.0744
L1	-0.1818	0.0675	-2.69	0.0071
P2	-0.1709	0.0392	-4.36	0.0000
P5	-0.0891	0.0474	-1.88	0.0604
M2	0.0985	0.0370	2.66	0.0078
M3	0.3104	0.1187	2.61	0.0089
M4	0.1749	0.0867	2.02	0.0437
M5	-0.1960	0.0811	-2.42	0.0157

Através de uma primeira análise à Tabela 4.20, o modelo apresenta valores baixos para o erro padrão (*Std. Error*) e através do *p-value* verifica-se que todos os fatores são

estatisticamente significativos.

Adicionalmente, é importante frisar que a variável (Intercept) refere-se ao Segurado Padrão, segurado este que será utilizado ao longo de toda a dissertação e que possui as seguintes características:

Tabela 4.21: Características do Segurado Padrão

Características	**Designação**	**Escalão**
Idade Condutor	[31,55]	IDA2
Localidade	Porto	Loc4
Anos de Carta	]16,30]	AC2
Marca	Mercedez Benz	M1
Tipo Veículo	Automóvel	G1
Combustível	Gasolina	C2
Utilização	Pessoal	T2
Lugares	5	L3
Idade do Veículo	[11,15]	IDAV3
Peso do Veículo	1.451-1.500 Kg.	P4
Cap. Cúbica	1.301-1.400 cm^3	CC3

Relativamente à Tabela 4.20 verifica-se que existem grupos de segurados que apresentam um risco mais elevado que o segurado padrão, como por exemplo, segurados com idades compreendidas entre o 17 e os 30 anos (IDA1) que tenha carta de condução à menos de 15 anos (AC1), que conduza um veículo da marca Ford (M2) da empresa (T1) a Diesel (C1). Estas resultados vão de encontro ao observado nas estatísticas preliminares, quando por exemplo se constatou que segurados mais novos apresentavam uma frequência de sinistralidade superior à frequência média da carteira.

Do mesmo modo, pode verificar-se que grupos de segurados apresentam um menor risco em relação ao segurado padrão. Segurados que residam por exemplo no distrito de Coimbra (Loc2) com mais de 75 anos de idade (IDA4), que conduza um veículo com mais de 30 anos (IDA5) da marca Jaguar (M5), com 4 lugares (L1), um peso bruto entre 1.401 e 1.450 quilogramas (P2), são um exemplo disso mesmo.

4.3 Modelação da Severidade

Encontrados os fatores e respetivos níveis tarifários que influenciam a frequência de sinistralidade, é necessário avaliar quais os fatores que influenciam a severidade dos sinistros.

Para a modelização desta variável aleatória, consideraram-se apenas as apólices que originaram sinistros, ou seja, 4.555 apólices às quais correspondem 4.860 sinistros.

O processo é semelhante ao adotado anteriormente, porém, é de frisar que nesta modelação não é necessário considerar a variável *offset*, uma vez que o custo associado ao sinistro não é influenciado pelo tempo de exposição ao risco.

Com o objetivo de estudar a distribuição do custo dos sinistros optou-se, em primeiro lugar, por efetuar uma análise à distribuição empírica dos dados, nomeadamente ao nível de simetria e existência de *outliers*, tal como se pode verificar na Figuras 4.16 e 4.17.

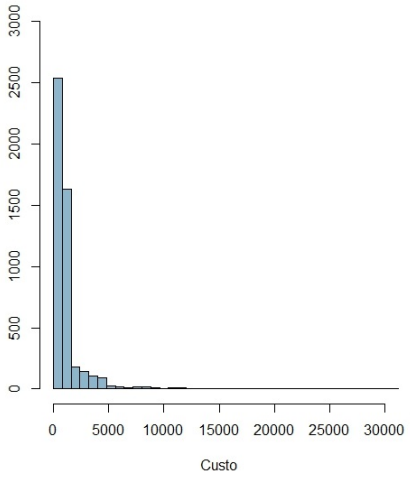

Figura 4.16: Histograma - Custo dos sinistros

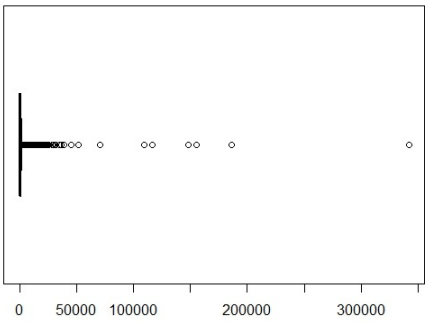

Figura 4.17: Boxplot - Custo dos sinistros

A análise do valores apresentados na Tabela 4.5 e a observação das Figuras 4.16 e 4.17 levou-nos a optar pela modelação em separada de sinistros de dimensão "usuais" e de "grandes sinistros", tal como referido na secção 2.2.1.3.

Analisando o gráfico da Figura 4.17, constata-se a existência de uma assimetria positiva, o que indica que a distribuição dos dados apresenta valores de mediana inferiores à média e também indica a existência de uma cauda à direita.

Para efeitos de construção da tarifa, e analisando os valores apresentados na Tabela 4.5, optou-se pela seguinte classificação:

- "Sinistros Usuais" - montante total do sinistro inferior a 12.000€;
- "Grandes Sinistros" - montante total do sinistro igual ou superior a 12.000€.

Faz-se notar que o montante de 12.000€ corresponde ao quantil de 99,92 % da distribuição empírica dos dados de sinistralidade.

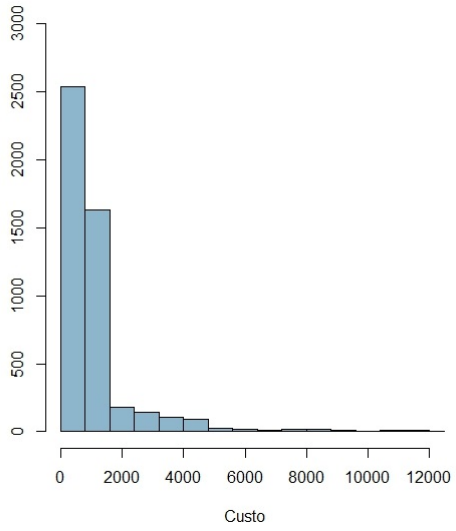

Figura 4.18: Histograma - Custo dos sinistros "usuais"

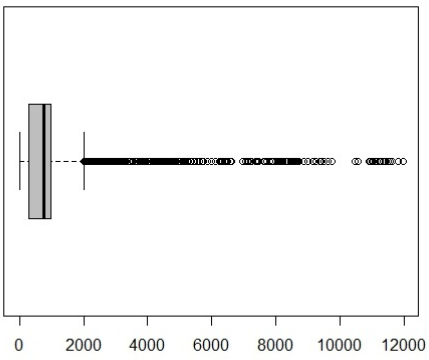

Figura 4.19: Boxplot - Custo dos sinistros "usuais"

Para obter o modelo que melhor represente a severidade dos sinistros, a primeira distribuição a ser considerada é a distribuição Gama, que segundo Brockman e Wright (1992) é uma distribuição que se ajusta bem a este tipo de dados.

As estimativas de máxima verosimilhança para os parâmetros da distribuição Gama, obtidas através do *software* R, são ilustrada na Tabela 4.22.

Tabela 4.22: Estimativas α e β - Gama

$\widehat{\alpha}$	$\widehat{\beta}$
0,981	1.046,121

Seguindo a metodologia de estudo adotada para a modelação da frequência de sinistralidade, na Tabela 4.23 são apresentadas as probabilidades de uma variável aleatória proveniente de uma família Gama(α,β) com os valores da Tabela 4.22. São também apresentadas as frequências esperadas e os valores do teste de ajustamento do Qui-Quadrado.

Tabela 4.23: Modelação da Severidade dos sinistros - Distribuição Gama

Montante	$\mathbf{O}_i$	$\mathbf{p}_i$	$\mathbf{E}_i$	χ^2_{calc}
]0-3.000]	4.432	0,945	4.549,420	3,03
]3.000-6.000]	293	0,051	248,765	7,87
>6.000	88	0,003	14,766	363,18
Total	4.813	1	4.813	374,08

O_i-N.º Sinistros p_i-Probabilidade E_i-Frequência Esperada

Analisando a Tabela 4.23, observa-se que o valor obtido para a estatística observada do teste, χ^2_{calc}, é de 374,08. Dado que este valor é superior a $\chi^2_{1;0,05} = 3,8411$, rejeita-se a hipótese de que os dados provenham de uma distribuição Gama de parâmetros α e β, com valores 0,981 e 1.046,121, respetivamente.

Rejeitada a hipótese anterior, uma distribuição alternativa a ser estudada é a Log-Normal. Suponha-se que $Z \sim$N(μ,σ^2). Com $Y = e^Z$ pode afirmar-se que a variável aleatória Y tem distribuição Log-Normal com parâmetros μ e σ^2. Reciprocamente, se Y tem distribuição Log-Normal, então $Z = \log(Y)$ tem distribuição N(μ,σ^2).

De modo a estimar os parâmetros da distribuição Normal, recorre-se ao método de Máxima Verosimilhança obtendo-se os seguintes valores:

Tabela 4.24: Estimativas μ e σ^2 - Distribuição Normal

$\widehat{\mu}$	$\widehat{\sigma^2}$
6,343	1,158

Seguindo a metodologia de estudo já adotada, na Tabela 4.25 são apresentadas as probabilidades de uma variável aleatória proveniente de uma distribuição Log-Normal com os respetivos parâmetros apresentados na Tabela 4.24.

Tabela 4.25: Modelação da Severidade dos sinistros - Distribuição Log-Normal

Montante	$\mathbf{O}_i$	$\mathbf{p}_i$	$\mathbf{E}_i$	χ^2_{calc}
]0-3.000]	4.432	0,924	4.448,97	0,06
]3.000-6.000]	293	0,054	262,88	3,45
]6.000-9.000]	62	0,012	59,83	0,07
> 9.000	26	0,004	20,84	1,27
Total	4.813	1	4.813	4,86

O_i-N.º Sinistros p_i-Probabilidade E_i-Frequência Esperada

Analisando a Tabela 4.25, observa-se que o valor obtido para χ^2_{calc} é de 4,86 ao qual corresponde um *p-value* de 2,74% pelo que, não se rejeita a hipótese de que os dados proveem de uma distribuição Log-Normal de parâmetros 6,343 e 1,158 para níveis de significância inferiores a 2,74%.

Recorrendo às mesmas funções utilizadas aquando da modelação do número de sinistros, efetua-se o estudo aos fatores e respetivos níveis tarifários que influenciam a severidade dos sinistros com o objetivo de obter um modelo simplificado, composto apenas por variáveis estatisticamente significativas. Para selecionar o modelo que apresente o melhor ajustamento, recorre-se novamente ao critério de seleção AIC e ao método *Stepwise Backward*.

Na tabela 4.26 são apresentados os valor do critério AIC para o modelo inicial e final, bem como o número de variáveis existente em cada modelo.

Tabela 4.26: Seleção de modelo - AIC

Modelo	N.º de variáveis	AIC
Modelo Inicial	30	15.070
Modelo Final	9	15.044

Ao analisar a tabela acima, verifica-se que o modelo final é composto apenas por 9 variáveis enquanto que o modelo inicial apresenta 30, bem como, o valor obtido pelo critério AIC é 15.070 e 15.044 para o modelo inicial e modelo final, respetivamente. Estes valores obtidos, indicam que o modelo final apresenta um melhor ajustamento comparativamente ao modelo inicial.

Encontrado o modelo que se ajusta aos dados da severidade dos sinistros, na Tabela 4.27 é apresentado o modelo final e as estimativas dos parâmetros de regressão, para cada nível tarifário bem como o respetivo *p-value*. Este é o modelo que será utilizado para a modelação da severidade dos sinistros.

Analisado a Tabela 4.27, os presentes fatores tarifários apresentam valores baixos para

Tabela 4.27: Estrutura Tarifária Final - Severidade dos sinistros

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2366	0.0358	174.28	0.0000
Loc6	0.1115	0.0512	2.18	0.0296
IDAV1/2	-0.0913	0.0387	-2.36	0.0184
IDAV4/5	0.0722	0.0425	1.70	0.0898
IDA3/4	0.0878	0.0512	1.72	0.0861
AC3/4	-0.1197	0.0526	-2.28	0.0229
M2	0.1107	0.0416	2.66	0.0078
CC1/2	0.1069	0.0354	3.02	0.0025
P5	0.1213	0.0519	2.34	0.0194
L2	0.1197	0.0434	2.76	0.0059

o erro padrão (*STd. Error*) e através da coluna Pr(>|t|), constata-se que todos os fatores são estatísticamente significativos.

Atendendo às estimativas dos coeficientes de regressão (coluna *Estimate*) presentes na Tabela 4.27, o perfil de risco que origina uma maior severidade de sinistros é o de um indivíduo com mais de 75 anos de idade (IDA3), que resida no distrito de Aveiro (Loc6), que conduza um veículo com mais de 30 anos (IDAV5) com 2 lugares (L2) da marca Ford (M2), que o veículo tenha um peso bruto entre os 1.551 e os 1.600 quilogramas (P5) e que tenha uma cilindrada entre os 1.601 e os 1700 cm^3 (CC1).

Relativamente ao perfil de risco do segurado que origina uma menor severidade em relação ao Segurado Padrão é, por exemplo, um individuo que conduza um veículo com idade entre os 6 e os 1 anos (IDAV2) e que tenha carta de condução à mais de 31 anos (AC3).

4.3.1 Tarifa MLG - Frequência/Severidade

Com ambos os modelos já definidos, é possível apresentar a estrutura tarifária final, com os respetivos prémios para a carteira automóvel em estudo. Na Tabela 4.28 são apresentados as estimativas da frequência de sinistralidade e da severidade dos sinistros para os respetivos escalões tarifários obtidos nos modelos das Tabelas 4.20 e 4.27. Por último é também apresentado o prémio relativamente ao Segurado Padrão e os respetivos descontos e agravamentos para os restantes escalões, considerando uma tarifa multiplicativa.

Os resultados da Tabela 4.28 incorporam a distribuição dos "grandes sinistros", seguindo a formulação referida na equação (2.9). As probabilidades $\mathbb{P}[Y < s|X]$ e $\mathbb{P}[Y \geq s|X]$ foram estimadas a partir da sinistralidade observada, sem serem consideradas as características observáveis de cada segurado, $\boldsymbol{X}$, efetuando uma distribuição uniforme por todos os escalões tarifários.

Tabela 4.28: Estrutura Tarifária Final Frequência/Severidade

Escalão Tarifário	**Frequência**	**Severidade**	**Prémio Puro**
Seg. Padrão	0,0986	1.237,88	156,51
Loc6	0,0818	1.383,91	94 %
Loc5	0,0910	**1.237,88**	94 %
Loc2/3	0,0671	**1.237,88**	75 %
AC1	0,1208	**1.237,88**	118 %
AC3/4	**0,0986**	1.098,18	91 %
IDA1	0,1160	**1.237,88**	114 %
IDA3/4	0,0836	1.351,54	94 %
IDAV1/2	0,0919	1.129,81	88 %
IDAV4/5	0,0905	1.330,55	99 %
T1	0,1633	**1.237,88**	151 %
C1	0,1155	**1.237,88**	113 %
P2	0,0831	**1.237,88**	88 %
P5	0,0902	1.397,56	103 %
M2	0,1088	1.382,76	118 %
M3	0,1344	**1.237,88**	128 %
M4	0,1174	**1.237,88**	115 %
M5	0,0810	**1.237,88**	86 %
L1	0,0822	**1.237,88**	87 %
L2	**0,0986**	1.395,34	110 %
CC1/2	**0,0986**	1.377,59	109 %

Observando os resultados obtidos na Tabela 4.28 verifica-se que os segurados que apresentam um perfil que originam um maior preço da tarifa são, por exemplo, indivíduos com menos de 30 anos de idade (IDA1), que residem no distrito de Porto (Loc4), que tenham carta de condução a menos de 15 anos (AC1), que conduzam um veículo da empresa (T1) a diesel (C1) de 2 lugares (L2) da marca Tata (M3), com um peso bruto compreendido entre os 1.551 e os 1.600 quilogramas (P5) e que possua um motor com cilindrada entre os 1.401 cm^3 e os 1.500 cm^3 (CC1).

Por outro lado, o perfil que origina uma menor tarifa são, por exemplo, os segurados com 56 anos de idade (IDA3), que residam em Coimbra (Loc2) com carta de condução à mais de 30 anos (AC3), que possuam um veículo com menos de 5 anos (IDAV1), com um motor a gasolina (C2), com uma capacidade cúbica entre os 1.101 e os 1.200 cm^3 (CC3) da marca Suzuki (M5) com 5 lugares (L3), com um peso entre os 1.351 e os 1.400 quilogramas (P2).

Na Tabela 4.29 são apresentados os prémios referentes ao Segurado Padrão bem como, o prémio máximo, ou seja, o prémio que sofreu mais agravamentos e o prémio mínimo, que refere-se ao prémio que obteve mais bonificações.

Tabela 4.29: Resumo de Prémios

Prémio Mínimo	Segurado Padrão	Prémio Máximo
58,79€	156,51€	564,59€

4.3.2 Tarifa - Família de Distribuição Tweedie

Era objetivo desta dissertação avaliar a possibilidade de obtenção de uma estrutura tarifária para a cobertura de responsabilidade civil automóvel com modelação efetuada através da Família de Distribuições Tweedie. Neste sentido, nesta secção, apresentam-se resultados obtidos para a estrutura tarifária.

Com o objetivo de obter a melhor estimativa para p, ver equação (3.5), e recorrendo ao código R disponibilizado por Charpentier (2014) obteve-se o valor de $\widehat{p} = 1,58$, que maximiza a função de verosimilhança da variável aleatória S, conforme apresentado na Figura 4.20.

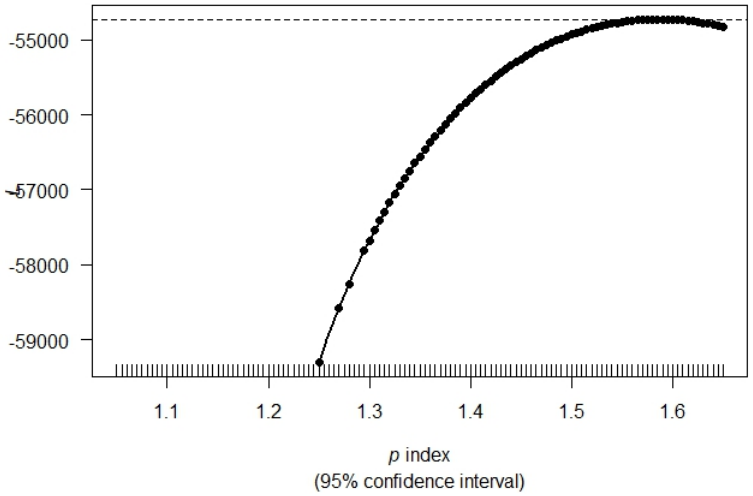

Figura 4.20: Função de verosimilhança vs p.

Observando o gráfico da Figura 4.20, constata-se que os valores obtidos para p vão de encontro aos resultados presentes na Tabela 3.2, em que para valores entre 1 e 2 faz referência ao Processo de Poisson Misto. Adicionalmente, é importante referir que para valores de p superiores a 1,65 o algoritmo utilizado não convergiu.

Estimado o valor de p e recorrendo ao *software* R obteve-se a estrutura tarifária para a cobertura de responsabilidade civil automóvel com base na Família de Distribuições Tweedie. Na seleção das variáveis a incorporar na tarifa foram efetuados os procedimentos descritos na secção 3.5.

Na Tabela 4.30 são apresentados os valores dos prémios obtidos para o segurado padrão bem como para os restantes escalões tarifários estatísticamente significativos para o modelo.

Tabela 4.30: Tarifa Final Tweedie

Escalão Tarifário	**Prémio Puro**
Seg. Padrão	115,18 €
Loc6	158,2%
AC1	212,7%
P2	66,5%

Com base nos resultados apresentados na Tabela 4.30 é possível determinar o perfil dos segurados que apresentam os prémios mais altos, e também os prémios mais baixos. Por exemplo, que resida em Aveiro (Loc6), com carta de condução à menos de 15 anos (AC1) origina um prémio com uma penalização de 236%. Por outro lado, um segurado que conduza um carro com um peso bruto entre os 1.351 e os 1.400 quilogramas (P2) tem uma bonificação de cerca de 66%.

Na Tabela 4.31 são apresentados os prémios puros referentes ao Segurado Padrão bem como, o prémio máximo e o prémio mínimo, referentes ao prémio que obteve respetivamente, mais penalizações e bonificações.

Tabela 4.31: Resumo de Prémios

Prémio Mínimo	Segurado Padrão	Prémio Máximo
76,60€	115,18€	387,53€

Analisando os prémios do segurado padrão em ambas as tarifas, pode-se constatar o prémio da tarifa estimada através dos Modelos Lineares Generalizados é superior ao prémio praticado quando utilizada a Família de distribuições Tweedie.

Quando comparados os prémios relativamente aos segurados que devam ser penalizados, conclui-se que a tarifa resultante da utilização da família Tweedie apresenta prémios mais baixos do que os prémios da primeira tarifa. Por outro lado, os prémios para "bons" segurados, ou seja, os prémios bonificados são maiores na tarifa quando utilizado o modelo Frequência/Severidade

Na Tabela 4.31 são apresentados exemplos de perfis de segurados e na Tabela 4.32 são apresentados os respetivos prémios estimados para cada estrutura tarifária.

Tabela 4.32: Exemplo de perfis de Segurados

Segurado	**S1**	**S2**	**S3**	**S4**
Idade Condutor	21	35	54	72
Localidade	Aveiro	Madeira	Lisboa	Bragança
Anos de Carta	3	16	31	52
Marca Veículo	Smart	Kawasaki	Volkswagen	Mercedez-Benz
Anos Veículo	8	2	10	24
N.º Lugares	2	1	7	5
Tipo Veículo	Carro	Motociclo	Carro	Carro
Combustível	Diesel	Gasolina	Gasolina	Gasolina
Utilização	Empresa	Pessoal	Pessoal	Pessoal
Cap.Cúbica	1.000 cm^3	1.200 cm^3	1.900 cm^3	1.500 cm^3
Peso	1.100 Kg.	200 Kg.	2.500 Kg.	1.900 Kg.

Tabela 4.33: Prémios Puros por Estrutura Tarifária

Segurado	**Estrutura Tarifária Freq./Sev.**	**Estrutura Tarifária Tweedie**
S1	331,76€	257,73€
S2	84,79€	115,18€
S3	112,59€	245,03€
S4	95,49€	115,18€

Analisando a Tabela 4.33, de um modo geral, na primeira tarifa, apesar de existir uma maior segmentação da carteira, esta apresenta prémios puros um pouco mais baixos, enquanto a segunda, apesar de ter uma menor segmentação, apresenta prémios puros mais elevados, que após introdução da carga de segurança e todas as restantes cargas inerentes aos contratos poderá produzir uma tarifa comercialmente exequível.

5

Conclusão

Com o aumento da competitividade do mercado segurador em Portugal, e em particular o crescimento registado nos últimos anos do ramo Não Vida, onde o ramo automóvel se tem vindo a tornar num ramo muito significativo para as contas das seguradoras, torna-se cada vez mais importante construir tarifas automóveis apelativas do ponto de vista comercial e rentáveis de modo a fazer face a futuras responsabilidades.

Na realização da presente dissertação, recorre-se à construção de uma tarifa *a priori*, abordagem esta que tenta prever a futura sinistralidade subdividindo a carteira em grupos de risco homogéneos. Para a construção da referida tarifa automóvel foram utilizados dois métodos, para os quais existem *packages* disponíveis no *software* R *Project*: o primeiro baseia-se na modelização em separado da frequência de sinistralidade e da severidade dos sinistros enquanto que o segundo método se centra na modelação das indemnizações agregadas.

É de realçar que apesar dos resultados obtidos no presente estudo, a utilização da família de distribuição Tweedie tem a vantagem de ser apenas necessário realizar um único estudo, sendo estimado desta forma o melhor valor para o parâmetro p.

Como estudos futuros, deveriam ser considerados outros métodos para a construção de tarifas automóveis e comparar o impacto destes com os métodos mais tradicionais. Exemplos a serem considerados são os Modelos Aditivos Generalizados (MAG) (ver Ohlsson e Johansson (2010)) e os Modelos Não Lineares Generalizados (MNG) (ver trabalho de Kaivanipour (2015) e Giorgio Giancaterino (2014)).

No decorrer da presente dissertação surgiram alguns obstáculos que dificultaram um desenvolvimento mais aprofundado. Exemplo disso foi a fase inicial de tratamento dos dados registados na base de dados que mostrou ser um processo moroso, pois existiam diversas incoerências e dados omissos que poderiam ter impacto significativo nos resultados finais e o caso da utilização da Família de Distribuição Tweedie quando se considera

que a severidade dos sinistros é modelada através da distribuição Log-Normal.

Considere-se que os resultados obtidos vão de encontro aos objetivos inicialmente propostos pela empresa Actuariado. Recorde-se que não fazia parte dos objetivos a proposta de uma tarifa comercialmente exequível, mas sim o estudo das técnicas que estão por base de projetos desta natureza.

Bibliografia

Achieng, O. M. (2010). "Actuarial modeling for insurance claim severity in motor comprehensive policy using industrial statistical distributions". Em: *International Congress of Actuaries, Cape Town*. Vol. 712.

Actuariado (www.actuariado.pt). *Actuariado - Estudos Actuariais, Económicos e Financeiros*. URL: www.actuariado.pt.

Akaike, H. (1974). "A new look at the statistical model identification". Em: *IEEE transactions on automatic control* 19(6), pp. 716–723.

Anderson, D., S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher e N. Thandi (2004). "A practitioner's guide to generalized linear models". Em: *Casualty Actuarial Society Discussion Paper Program*, pp. 1–116.

ASF (2016). *Autoridade de Supervisão de Seguros e Fundos de Pensões "O presente e o futuro da atividade seguradora em Portugal"*.

Brockman, M. J. e T. Wright (1992). "Statistical motor rating: making effective use of your data". Em: *Journal of the Institute of Actuaries* 119(3), pp. 457–543.

Burnham, K. P. e D. R. Anderson (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Centeno, M. (2002). *Teoria do Risco na Actividade Seguradora*. Celta Editora.

Charpentier, A. (2014). *Computational Actuarial Science with R*. Chapman e Hall/CRC.

Clark, D. R. e C. A. Thayer (2004). "A primer on the exponential family of distributions". Em: *Casualty Actuarial Society Spring Forum*, pp. 117–148.

Cox, D. R. e D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

De Jong, P. e G. Z. Heller (2008). *Generalized linear models for insurance data*. Vol. 10. Cambridge University Press Cambridge.

Fahrmeir, L. e H. Kaufmann (1985). "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models". Em: *The Annals of Statistics*, pp. 342–368.

Fahrmeir, L. e G. Tutz (2001). "Models for multicategorical responses: Multivariate extensions of generalized linear models". Em: *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, pp. 69–137.

Frees, E. W., G. Lee e L. Yang (2016a). "Multivariate frequency-severity regression models in insurance". Em: *Risks* 4(1), p. 4.

Frees, E. W., R. A. Derrig e G. Meyers (2016b). *Predictive modeling applications in actuarial science*. Vol. 2. Cambridge University Press.

Giorgio Giancaterino, C. (2014). *GLM, GNM and GAM approaches on TPML pricing*.

Goovaerts, M, R Kaas, J Dhaene e M Denuit (2001). *Modern actuarial risk theory*.

Grandell, J. (1997). *Mixed poisson processes*. Vol. 77. CRC Press.

Guerreiro, G. (2016). *Manual de Construção de Tarifas com R - O Exemplo do Seguro Automóvel*. FCT NOVA.

Hosmer Jr, D. W., S. Lemeshow e R. X. Sturdivant (2013). *Applied logistic regression*. Vol. 398. John Wiley & Sons.

Jorgensen, B. (1997). *The theory of dispersion models*. CRC Press.

Jørgensen, B. e M. C. Paes De Souza (1994). "Fitting Tweedie's compound Poisson model to insurance claims data". Em: *Scandinavian Actuarial Journal* 1994(1), pp. 69–93.

Kaas, R., M. Goovaerts, J. Dhaene e M. Denuit (2008). *Modern actuarial risk theory: using R*. Vol. 128. Springer Science & Business Media.

Kaivanipour, K. (2015). *Non-Life Insurance Pricing Using the Generalized Additive Model, Smoothing Splines and L-Curves*.

Lemaire, J. (1995). *Automobile insurance: actuarial models*. Springer Science & Business Media.

McCullagh, P. (1984). "Generalized linear models". Em: *European Journal of Operational Research* 16(3), pp. 285–292.

Nelder, J. e R. Wedderburn (1972). "General linearized models". Em: *J. Roy. Stat. Soc. Ser. A* 135, pp. 370–384.

Ohlsson, E. e B. Johansson (2010). *Non-life insurance pricing with generalized linear models*. Vol. 21. Springer.

Rao, C. R. (2009). *Linear statistical inference and its applications*. Vol. 22. John Wiley & Sons.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Ross, S. (1996). *Stochastic processes*. Wiley series in probability and statistics: Probability and statistics. Wiley. ISBN: 9780471120629.

Schwarz, G. (1978). "Estimating the dimension of a model". Em: *The annals of statistics* 6(2), pp. 461–464.

Sen, P. K. e J. M. Singer (1994). *Large sample methods in statistics: an introduction with applications*. Vol. 25. CRC Press.

Snedecor, G e W Cochran (1989). *Stadistical methods. Eight Ed. Ames*.

Turkman, M. A. A. e G. L. Silva (2000). "Modelos Lineares Generalizados-da teoria à prática". Em: *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*.

Tweedie, M. (1984). "An index which distinguishes between some important exponential families". Em: *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*. Vol. 579, 604.

Wilks, S. S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses". Em: *The Annals of Mathematical Statistics* 9(1), pp. 60–62.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

A

Anexo A

Tabela A.1: Descrição da variável *PesoVeiculo*

Escalão de Peso	Descrição
001	1-250 Kg.
002	251-500 Kg.
003	501-750 Kg.
004	751-1.000 Kg.
005	1.001-1.250 Kg.
006	1.251-1.300 Kg.
007	1.301-1.350 Kg.
008	1.351-1.400 Kg.
009	1.401-1.450 Kg.
010	1.451-1.500 Kg.
011	1.501-1.550 Kg.
012	1.551-1.600 Kg.
013	1.601-1.650 Kg.
014	1.651-1.700 Kg.
015	1.701-1.750 Kg.
016	1.751-1.800 Kg.
017	1.801-1.850 Kg.
018	1.851-1.900 Kg.
019	1.901-1.950 Kg.
020	1.951-2.000 Kg.
021	2.001-2.250 Kg.
022	2.250-2.500 Kg.
023	2.501-2.750 Kg.
024	2.751-3.000 Kg.
025	3.001-3.250 Kg.
026	3.250-3.500 Kg.
027	≥3501 Kg.

Tabela A.2: Descrição da variável *CapCubVeiculo*

Escalão de Capacidade Cúbica	Descrição
001	1-250 cm^3
002	251-500 cm^3
003	501-750 cm^3
004	751-1000 cm^3
005	751-1.000 cm^3
006	1.001-1.100 cm^3
007	1.101-1.200 cm^3
008	1.201-1.300 cm^3
009	1.301-1.400 cm^3
010	1.401-1.500 cm^3
011	1.501-1.600 cm^3
012	1.601-1.700 cm^3
013	1.701-1.800 cm^3
014	1.801-1.900 cm^3
015	1.901-2.000 cm^3
016	2.001-2.100 cm^3
017	2.101-2.200 cm^3
018	2.201-2.300 cm^3
019	2.301-2.400 cm^3
020	2.401-2.500 cm^3
021	2.501-3.000 cm^3
022	3.001-5.000 cm^3
023	$\geq$10.000 cm^3

Tabela A.3: Descrição da variável *Marcas*

Escalão	**M1**	**M2**	**M3**	**M4**
Descrição	AUDI	CITROEN	BEDFORD	LANCIA
	BMW	FORD	CHEVROLET	LAND ROVER
	FIAT	PEUGEOT	DACIA	SMART
	HONDA		DAEWOO	VOLVO
	HYUNDAI		DAIHATSU	
	MERCEDES BENZ		DUCATI	
	MITSUBISHI		GALLOPER	
	NISSAN		ISUZU	
	OPEL		IVECO	
	RENAULT		JEEP	
	ROVER		KTM	
	SEAT		LEXUS	
	TOYOTA		LOTUS	
	VOLKSWAGEN		TATA	
Escalão	**M5**			
Descrição	AJS	DODGE	LADA	PORTARO
	ALFA ROMEO	DUTTON	LONCIN	RANGE ROVER
	ALPINE	EBRO	MALAGUTTI	RIMOR
	ASTON MARTIN	ELNAGH	MAN	ROLLS ROYCE
	AUSTIN	FERRARI	MASERATI	SAAB
	AUTOBIANCHI	GRANDIN	MATRA	SADO
	AUTOSTAR	GTM	MAZDA	SANCHS
	BENIMAR	HARLEY DAVIDSON	METRO	SIS-SACHS
	BERTONE	HERSTELLER	MG	SKODA
	BURSTNER	HUMMER	MINI	SSANGYONG
	CASAL	HUSQVARNA	MINI MOKE	SUBARU
	CATERHAM	HYMER	MOBILVETTA DESIGN	SUZUKI
	CHALLENGER	HYOSUNG	MONCAYO	SYM
	CHRYSLER	JAGUAR	MORGAN	TALBOT
	CONFERSIL	KAISER	MORRIS	TRIUMPH
	DAF	KAWASAKI	PANHARD	UMM
	DAFIER	KEEWAY	PIAGGIO	VESPA
	DAIMLER	KIA	PIAGGIO-VESPA	WILLYS
	DATSUN	KIMCO	PONTIAC	YAMAHA
	DERBI	KYMCO	PORSCHE	ZUNDAPP

B

Anexo B

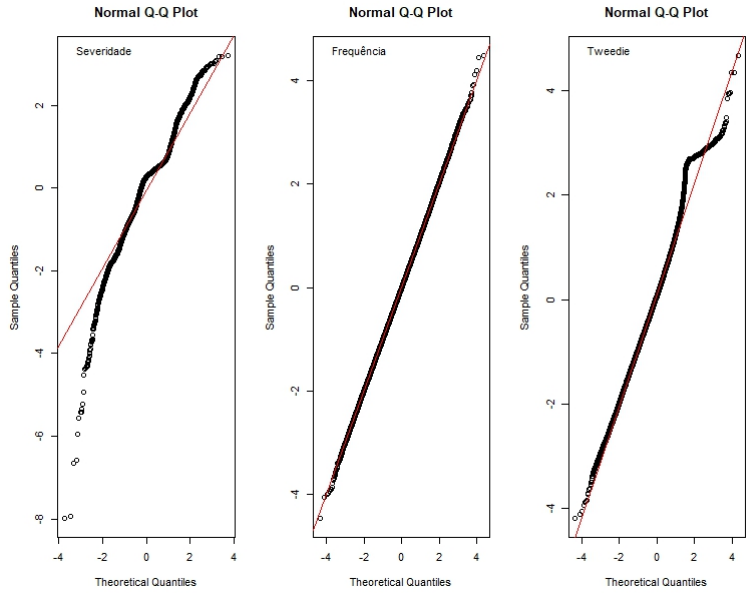

Figura B.1: Ajustamento da Frequência, Severidade e Tweedie